Characterizing and Recognizing Spoken Corrections in Human-Computer Dialog

by

Gina-Anne Levow

B.A./B.A.S, University of Pennsylvania (1989)
S.M., Massachusetts Institute of Technology (1993)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1998

Signature of Author .. ..
Department of Electrical Engineering and Computer Science
September 4, 1998

Certified by..... .....
Robert C. Berwick
Professor of Computer Science and Engineering and Computational Linguistics
Thesis Supervisor

Accepted by..... .....
Arthur C. Smith
Chairman, EECS Committee on Graduate Students

Characterizing and Recognizing Spoken Corrections in Human-Computer Dialog

by

Gina-Anne Levow

Revised version of a thesis submitted to the
Department of Electrical Engineering and Computer Science
on September 4, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Miscommunication in human-computer spoken language systems is unavoidable. Recognition failures on the part of the system necessitate frequent correction attempts by the user. Unfortunately and counterintuitively, users' attempts to speak more clearly in the face of recognition errors actually lead to decreased recognition accuracy. The difficulty of correcting these errors, in turn, leads to user frustration and poor assessments of system quality.

Most current approaches to identifying corrections rely on detecting violations of task or belief models that are ineffective where such constraints are weak and recognition results inaccurate or unavailable. In contrast, the approach pursued in this thesis, in contrast, uses the acoustic contrasts between original inputs and repeat corrections to identify corrections in a more content- and context-independent fashion.

This thesis quantifies and builds upon the observation that suprasegmental features, such as duration, pause, and pitch, play a crucial role in distinguishing corrections from other forms of input to spoken language systems. These features can also be used to identify spoken corrections and explain reductions in recognition accuracy for these utterances. By providing a detailed characterization of acoustic-prosodic changes in corrections relative to original inputs in a voice-only system, this thesis contributes to natural language processing and spoken language understanding. We present a treatment of systematic acoustic variability in speech recognizer input as a source of new information, to interpret the speaker's corrective intent, rather than simply as noise or user error. We demonstrate the application of a machine-learning technique, decision trees, for identifying spoken corrections and achieve accuracy rates close to human levels of performance for corrections of misrecognition errors, using acoustic-prosodic information. This process is simple and local and depends neither on perfect transcription of the recognition string nor complex reasoning based on the full conversation. We further extend the conventional analysis of speaking styles beyond a 'read' versus 'conversational' contrast to extreme clear speech, describing divergence from phonological and durational models for words in this style.

Thesis Supervisor: Robert C. Berwick
Title: Professor of Computer Science and Engineering and Computational Linguistics

Acknowledgements

Although my name appears on this thesis, I could never have completed it without the support and assistance of many others, too numerous to name here. I will, however, take a few lines to thank those whom I owe special debts: To Nicole Yankelovich and the Sun Microsystems Labs Speech Group for the use and transcription of the data from the SpeechActs field trial. To Sharon Oviatt and the Center for Human-Computer Communication for focusing my efforts on spoken corrections and providing the structure and tools for the acoustic analyses in this work. To my thesis supervisor, Robert Berwick, for encouragement, guidance, and the freedom to pursue this area of research. To the members of my thesis committee, Justine Cassell and Paul Viola, for insightful feedback and a commitment to high standards of research. To Eric Grimson for invaluable aid traversing the academic and administrative pitfalls of Area II. To my fellow students and lab members for a uniquely MIT combination of unflinching criticism, unstinting support, and wide-ranging discussion: Robbin Chapman, Carl de Marcken, Charles Isbell, Tina Kapur, Marina Meila, Latanya Sweeney, Holly Yanco, Charles Yang, Liana Lorigo, and many, many others. To my family, for their love, support, and absolute conviction that this could be done. And, last and most important, to Jim, for everything.

Contents

Chapter 1

Introduction

Recognition errors are an unavoidable problem for today's recognition- based interfaces. While isolated word speech recognition error rates can be as low as 5%, a recognizer working on continuous, conversational speech input often reaches

25% error for recognition of full utterances, due to the increased complexity of this task. These frequent recognition errors naturally necessitate frequent corrections. The severity of these problems can be demonstrated by a simple example of an error "spiral", a sequence of repeated recognition errors on the part of the system and repeated attempts to enter the same input on the part of the user, as shown below.

USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8 since yesterday.
USER: Give me the price for AT&T.
SYSTEM: Data General is at 63 1/4, up 1/3.
USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8.
USER: Give me the price for American Telephone and Telegraph.
SYSTEM: Excuse me?

These sequences are all too common in speech recognition systems and are naturally a source of great frustration to users. Each of the user inputs that follows a recognition error attempting the original misrecognized input is considered a correction. Understanding how corrections differ from other inputs and how these differences can be used to recognize such acts will form the focus of this thesis.

It has been observed [Shriberg *et al.*, 1992], [Rudnicky and Hauptmann, 1990] that recognizer error rates are higher for corrections than for other input. This phenomenon suggests two important possibilities: first, that this observed increase in recognition errors results from characteristics of correction utterances which differ from other inputs and the recognizer's model of expected input, and secondly, that these changes in speaking style in correction utterances contribute to the frustrating phenomenon of error spirals, such as the one shown above.

Another important reason to more carefully examine error corrections comes from user assessments of system quality. In a recent study [Sukkar *et al.*, 1996] researchers asked users to evaluate the quality of several spoken language systems. The study

revealed that, even for recognition system with low error rates, users gave very unsatisfactory reviews of systems where it was more difficult to recover from errors when they did arise. These issues of decreased recognition accuracy, contribution of error spirals, and impact on system usability strongly argue on practical grounds that we focus our attention on a more thorough understanding of corrections of recognition errors in spoken language systems.

In addition to these very practical arguments, it is also important to better understand the process of corrections in order to identify not only the lexical content of the user's input, but also the intent with which the input was made. In other words, we would like to understand the real meaning of the utterance, at the level of the speech act the user performed. Understanding that an utterance represents a correction of previous input, rather than new input has direct ramifications on the actions that should be performed by the system. Consider the following scenario as an example. The system has just received an input that it interprets as "Delete message two." The user then attempts a correction, saying "Delete message EIGHT." An appropriate response to this input as a new input would simply be to delete the eighth message. However, an appropriate response to this input when interpreted as a correction would be to delete the eighth message and, *in addition*, reverse or offer to reverse the deletion of message number two, in accord with the user's corrective intent. Thus, identifying the corrective intent of an utterance can have an important impact on the choice of appropriate system action in response to that utterance.

1.1 Challenges

Unfortunately, identifying an utterance in human-computer dialog as a correction is far from simple. With a full accurate transcription of both user input and system response as shown in the initial error spiral in this chapter the user's repeated attempts at correcting the system are painfully obvious. However, the problem of error spirals arises form the simple fact that the system often does not have an accurate transcription of user input. Specifically, the system either misrecognizes the user's input, giving an erroneous transcription as seen in the first three responses in the error spiral, or rejects the input outright, failing to obtain any adequate result from the recognizer as occurs in the last step shown in the spiral. Thus the system does not have any guarantee of the ability to compare successive user inputs accurately; if it did, the need for correction would not have arisen in the first place. Many strategies for detecting self-repairs, where the user corrects their own utterance in mid-stream use lexical similarity between the original section of the input and the corrected component to detect the repair action, as we will discuss in more detail in the chapter on related work. Leaving aside for the moment the difficulty of obtaining a text transcription in the full speech recognition environment, the simple act of repeating a command to a spoken language system does not necessarily imply an attempted correction. An examination of 7752 user utterances in SpeechActs, a spoken language system, revealed that approximately 500 distinct phrases constituted almost 6700 of the observed utterances. In other words, a mere 500 phrases accounted for 80% of

the inputs to the system. Approximately, 1000 text strings occurred only once. Many utterances do appear many times, without necessarily involving any corrective intent. For example, many commands in these systems involve navigation through lists of information available to the user. It is not unusual for a user to simply navigate through the list by repeatedly entering the command "next" or "next message". (One subject, in particular, often completed entire session with the system by simply logging in , saying "next message" or "skip it" through all the messages, and hanging up.) However, by far the most common strategy employed by users in making a correction was the simple repetition of the same lexical content as the original input attempt.

If similarity of lexical content is not sufficient to identify spoken corrections, perhaps there are specific lexical cues to the discourse function of these utterances. A number of cue words or phrases are known to signal different discourse functions such as topic shift, acknowledgment, or explanation. [Reichman, 1985], [Hirschberg and Litman, 1993]. There are even cue phrases which are associated with corrections, such as a sentence-initial "no" or "I meant". However, ironically, these cues were found only rarely in transcripts of user interactions with a spoken language system. Only seven utterances of over 7700 contained such cues; a similar number of profanities were encountered. Clearly, one can not rely upon the presence of lexical cues to signal corrections in human-computer dialog.

The preceding discussion shows that lexical information alone is not sufficient to identify an utterance as a correction, nor would lexical usage effectively explain the greater frequency of recognition errors observed on correction utterances. Clearly, however, there is something distinctive about correction utterances that allows human listeners to identify corrections, even in isolation, at almost 80% accuracy. This identification argues that, not lexical, but suprasegmental features often signal the corrective intent of an utterance.

The use of suprasegmental variation, such as changes in duration, pause, or pitch, to distinguish corrections from other utterances would explain, not only how people could identify corrections even in isolation, but also the dual problems of the difficulty of recognizing correction in a common speech recognizer context as well as the lower recognition accuracy observed on correction utterances. First, features such as fundamental frequency (pitch) and amplitude (loudness) that are, at least anecdotally, associated with spoken corrections to computers and experimentally linked with corrections to speakers in other populations, such as the hard-of-hearing or children, are generally stripped off or normalized away in most current speech recognition systems. Since this information is removed, it would be inaccessible for the purposes of identifying corrections. Other common suprasegmental features, such as change in duration or pause, would, in fact, present direct difficulties for speech recognizers. This difficulty arises because speech recognition relies upon a match between the recognizer model and the observed input in durations, since an explicit penalty is imposed on recognition hypotheses in which phoneme durations exceed expected model durations.

This thesis quantifies and builds upon this observation that suprasegmental features, such as duration, pause, and pitch, play a crucial role in distinguishing corrections from other forms of input to spoken language systems and that the features can,

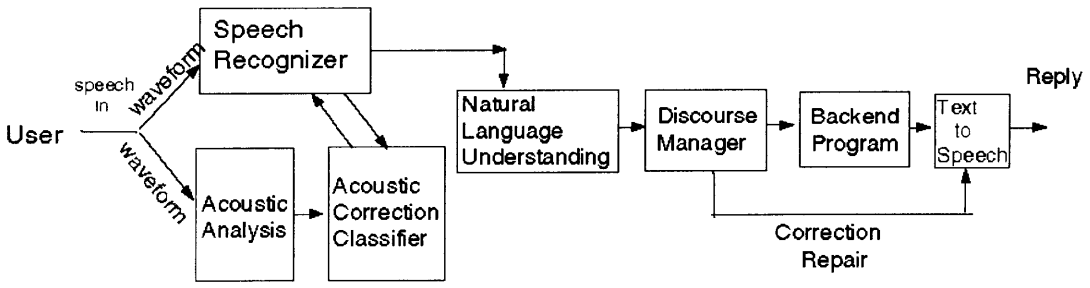


Figure 1-1: Architecture of System with Correction Classifier

in addition, be used to identify spoken corrections and explain reductions in recognition accuracy. By providing a detailed characterization of acoustic-prosodic changes in correction utterances relative to original inputs in human-computer dialog with a proto-type fully voice-in/voice-out spoken language system, this thesis contributes to natural language processing and spoken language understanding. We present a treatment of systematic acoustic variability in speech recognizer input as a source of new information, to interpret the speaker's corrective intent, rather than simply as noise to be normalized or a bad habit that the user should mend. We demonstrate the application of a machine-learning technique, decision trees, and achieve accuracy rates close to human levels of performance for corrections of misrecognition errors, using acoustic-prosodic information to identify spoken corrections. This process is simple and local and depends neither on perfect transcription of the recognition string or complex reasoning based on the full conversation. We further extend the conventional analysis of speaking styles beyond a read versus conversational contrast to extreme hyper-clear speech, describing divergence from phonological and durational models for words in hyper-clear speech.

1.2 System Design

The classifier would be incorporated into a standard spoken language system architecture as depicted above. (Figure 1-1) First, all user utterances would be passed, in parallel, to both the base speech recognizer and a speech act (in this case, correction) classifier module. The first component of this classifier module performs simple acoustic analysis: utterance start and end checkpointing for duration analysis, pause detection, pitch and amplitude tracking, and speaking rate computation. These measures are then fed into the correction classifier itself. The classifier will identify the utterance as corrective or not based on these measures. If a recognition result becomes available from the speech recognizer (SR) unit, the classifier can incorporate that information as well. If the utterance is classified as non-corrective, the recognition result is passed, as usual, through the remainder of the system unimpeded. However, if the utterance is classified as a correction, two steps follow. First, this information is sent back to the SR module to invoke any acoustic adaptation rules for correction recognition and the utterance is reanalyzed. Second, a message is passed to

the discourse manager signaling that a correction has been detected. This discourse manager will then intervene in the processing of the utterance.

There are a number of forms this intervention could take. First the previous recognition string, stored in the discourse history stack, is marked as incorrect. If available, the current and previous recognition results are lexically compared and single points of substitution are identified. If a single field differs, the system can prompt the user to explicitly enter that field, confirm it, or shift entry styles and ask the user to spell the item or select it from a list, for instance. The system can, even without testing for specific mismatch, present an explicit help message to the user, indicating recognition trouble or simply shift to a more directive interactive style, improving on the simple method, employed in many systems, of using longer prompts after repeated rejections.

We can follow this process through to see how it could alleviate the problems encountered in our example error spiral. After the system misrecognizes the first attempt to get the price for AT&T, the user attempts a correction. The acoustic analysis identifies an increase in utterance duration and pause duration. The classifier recognizes this change as signaling a correction. It marks the recognition of 'Hewlett-Packard' in the first input as an error and adjusts the context to remove reference to 'H-P' as the current company. It then performs a lexical comparison of the original and correction recognition results, identifying the company field as suspect. The system then shifts to directive mode, asking the user for the name of the company to check. The user would then enter 'AT&T'. The system could then prompt for explicit confirmation. Detecting the corrective intent of the user thus can defuse error spirals before the user becomes frustrated.

1.3 Preview

This thesis analyzes correction utterances in comparison to basic original inputs to develop a more precise characterization of the acoustic- prosodic differences between these two classes of utterances. This characterization in turn informs the design of a classifier to distinguish utterances of these two classes and finally can suggest modifications of speech recognizer design to improve recognition accuracy in the face of correction-related speech adaptations.

We begin with a discussion of related research. While little work has directly focussed on corrections, other work in discourse and dialog understanding can inform the analysis of corrections. Corrections can be viewed as performing a specific function in a discourse, as initiating a correction subdialog or performing a corrective speech act. Thus we look at research into the acoustic correlates of discourse structure and of different types of discourse relations. Here research indicates that utterance duration and pitch range expand when a new subdialog is initiated and that in some cases specific pitch contours may be associated with specific discourse relations. Corrections can also be viewed as a kind of "repair" of a failure of the "other" participant in the dialog. In this view, we examine work on other dialog repairs, specifically self-repairs, where the speaker corrects themselves, looking at lexical and acoustic cues to

repairs. Here we find that repetition and silence within the utterance and increases in duration, pitch, and amplitude can identify the location of a self-repair. Finally, we look at two purely descriptive analyses of spoken correction, in systems with only textual feedback, where increases in utterance duration and pause length characterize corrections.

We then set the stage for a more detailed analysis of corrections by describing the spoken language system, SpeechActs, through which we collect the data for these experiments. The system is a voice-only interface and thus relies exclusively on acoustic cues, differing from other platforms where corrections have been studied. We find that users interacting with the system encounter an overall recognition error rate of 25%, over more than 60 hours of recorded interactions. These errors are divided between two-thirds *rejection* errors, complete recognition failures, and one-third *misrecognition* errors, where a recognition result is found but contains some mistakes. We explore the source of these errors in greater depth, beginning with a comparison of error rates for novice and expert users. We find that novices improve their error rates over time, largely by learning the vocabulary and constructions understood by the system. However, even without these vocabulary related errors, novice users still encounter more errors than experts. The almost complete elimination of out-of-vocabulary utterances from novices' interactions can be tied to a decrease in their working vocabularies over time, to a set of words in which the user has high confidence. We also find that although each user converges on a small working vocabulary, there is still large variation in vocabulary between users. The presence, however, of persistent errors, an average of almost 20% for even more experienced novice users, illustrates the importance of properly handling errors and corrections.

Now we move on to the specific acoustic analysis of original inputs in contrast to repeat corrections. We use a data set consisting of approximately 300 lexically matched original-repeat utterance pairs. The "original" represents the user's first attempt at a given command, and the repeat correction corresponds to the first retry after the spoken recognition error. We analyze these utterances across a range of acoustic prosodic measures.

Broad Class	Specific Measures
Duration	Total utterance duration, speech duration
Pause	Total and Average Pause duration
	Proportion of Silence
Amplitude	Average and Maximum Loudness
Pitch	Maximum, Minimum, Range
	Internal contour slope, final contour

We find significant differences between original inputs and repeat corrections for measures of all types except amplitude. There are significant increases in all duration and pause measures. Pitch minima are lower and final contours are more frequently falling, for corrections in contrast to original inputs. Finally, for corrections of misrecognition errors only, utterance-internal contours are more variable, having steeper rises and larger cumulative slopes.

These contrasts demonstrate a large difference between original inputs and repeat corrections and suggest a set of features to be used in classifying utterances as originals or corrections. We design decision tree classifiers to perform this task, exploiting their ability to ignore irrelevant attributes and to produce readily intelligible and interpretable classifiers. We develop a set of 38 features for this classifier incorporating the measures found to be significant in the acoustic analysis, such as duration, pause, and pitch. These measures are used in both absolute and normalized forms, since the absolute values are highly variable. For instance, duration ranges from 210 to 5130 milliseconds, depending mostly of lexical content rather than original/correction status. We take as our basic unit of analysis a single user utterance, defined as the region between the system's prompt tone for user input and the system's next response, with preceding and trailing silences clipped. We emphasize identifying isolated individual utterances as original or corrective, but also examine the possibility of improving classification by comparing *pairs* of utterances for contrasts characteristic of corrections. We build classifiers that achieve between 65-77% accuracy, relative to a 50% baseline, depending on the type of correction being classified and the amount of information, full text transcription versus purely acoustic speaking rate measures. We find absolute and normalized duration to be the most important measures, producing the first split in all trees. For all correction types, other important features are pause duration and proportion of silence in the utterance. For corrections of misrecognition errors only, pitch contour and pitch minimum play an important secondary role. When compared to a 79.4% accuracy rate for human subjects on a similar task, the accuracy rates for classifiers fare well.

Finally, we consider the ramifications of the large differences between original inputs and repeat corrections for the design of adaptive speech recognizers. First we observe that there are a number of phonological contrasts between originals and corrections. These shifts are largely a natural effect of the increased duration and pause lengths in correction utterances. The majority of these changes involve a shift to a clearer, more careful speaking style in corrections, and away from a more conversational, casual style in original inputs. The contrasts include shifts from reduced vowels and consonants, such as 'schwa' or flapped 't', to unreduced, citation forms, such as a full vowel 'oo' in 'to' or a released, aspirated 't'. In addition we observed shifts to what can be called a "hyper-clear" style, characterized by insertion of a vowel or syllable, in correction utterances, as in shifting from 'goodbye' to 'good-ba-aye'. All of these changes constitute a shift away from the model of pronunciation derived from a base lexicon with standard co-articulation modeling. When we look at durational contrasts with a basic speech recognizer, rather than original inputs, the contrasts are even more clear. We find that a basic durational model, derived from TIMIT utterances [Chung, 1997], provides a good fit for words in non-final position in original inputs, with an almost normal distribution centered slightly above the TIMIT mean. Words in final position, though, begin to diverge from the model, being 0.5 standard deviations longer than predicted. However, corrections present a very poor match with this model, but in a systematic fashion. Durations of both final and non-final words become longer, moving the distributions 0.25 to 0.5 standard deviations *further* from the mean. We can thus see that a speech recognizer would need to take into

account these discourse and sentence-level information sources in order to adapt to recognizing corrections, particularly in those utterance- final words where correction and phrase-final lengthening effects combine.

In conclusion, we discuss specific ways to incorporate this understanding of durational change in corrections to build a more context- adaptive speech recognizer, either by incorporating phonological change rules conditioned on the corrective status of the utterance, or by modifying the durational model or scoring penalty for corrections. We also discuss the application of a similar decision tree method to isolate the corrected word(s) in the correction of misrecognition errors. Finally, we explore the possibility of using the original-repeat *pair* adjacency information to build a pair-based, rather than isolated correction recognizer.

1.4 Outline

The remainder of the thesis will follow the outline below. We will begin in Chapter 2 with an examination of related research. We will demonstrate how this thesis draws upon work in the use of intonation to identify discourse structure and discourse function. We will also look at research into recognizing the presence and position of self-repairs, another form of correction where the user corrects him/herself rather than the system. We will describe the use of both lexical and acoustic-prosodic information to solve this problem. Finally, we will discuss the small body of work explicitly involved corrections in human-computer dialog.

Chapter 3 will present a detailed description of SpeechActs, the prototype spoken language system used to collect the human-computer dialog data on which this thesis is based. We will describe typical interactions with the system and the subject population. We will then take a high-level look at error rates and error types encounter by users and examine in more depth the differences between novice and expert users in terms of error rates, use of out-of-vocabulary utterances, and vocabulary size and rate of acquisition. We will explain the selection of the original input-repeat correction pair data used extensively in acoustic analysis and classification experiments that form the core of the thesis.

Chapter 4 provides a detailed characterization of the acoustic-prosodic changes in correction utterances relative to original inputs in human-computer dialog. We examine four classes of acoustic-prosodic measures: duration, pause, pitch, and amplitude. We explain the significant differences between original inputs and repeat corrections for three of these classes: duration, pause, and pitch, and identify important contrasts between corrections of rejection errors and corrections of misrecognition errors. Finally, we relate these contrasts to a continuum in speaking style from conversational to clear.

Chapter 5 demonstrates the use of the acoustic features analyzed in the previous chapter in the development of a machine-learning based classifier to distinguish between corrections and other inputs. We begin by explaining the choice of decision trees as the classifier mechanism, for reasons for intelligibility and robustness to irrelevant attributes. We then describe in detail the feature set used to construct the

decision tree classifiers. We present results for decision tree classifiers on different correction types and using different types of lexical and contextual information.

Chapter 6 explains how the contrasts between original inputs and repeat corrections reflect a shift away from the models expected to support the speech recognizer. We demonstrate phonological contrasts between repeat utterances and both conversational and clear speech models of pronunciation. We also highlight the contrasts between the observed durations of correction utterances and durations predicted by a typical speech recognizer model, providing suggestions for modifications to a speech recognizer that would make it more effective in the face of correction-related adaptations.

Chapter 7 summarizes the results from the thesis and presents several paths for future work. We discuss accommodations in the speech recognizer to improve recognition of corrections. We describe the application of acoustic features related to corrections to the identification of the word or segment being corrected and present some preliminary results in this area.

Chapter 2

Related Work

Fully operational spoken language systems are a very recent development. Consequently, there has been relatively little experimental analysis of users' interactions with such systems, much less analysis of detailed correction interactions in such systems. We draw on related work in three main areas to provide direction and comparison for this work on characterizing and recognizing spoken corrections. First we consider work done in the broader area of understanding the structure of extended narrative or dialog sequences, usually referred to as discourse. Understanding the structure of discourse involves both identifying shifts in topic and relationships between topics and utterances within topics. Such relationships include correction, acknowledgement, and clarification, for instance. We will focus of identifying these structures in spoken interactions, emphasizing work on using prosodic cues, such as duration, loudness, and pitch. Second, we discuss the body of work most closely linked to corrections, that of automatic recognition of self-repairs. Self-repairs arise when a speaker interrupts themself to correct all or part of the current utterance; the later part of the utterance corrects an earlier portion. Corrections, as studied in this thesis, arise when the speaker's current utterance is intended to correct a perceived misunderstanding, misrecognition, or recognition failure of the part of the *other* conversant, and thus form a natural parallel to self-repairs. Finally we will examine the small body of work that has directly studied spoken corrections, highlighting the different environments and tasks in which the data was collected.

2.1 Inferring Discourse Relations from Speech

2.1.1 Inferring Discourse Structure from Speech

In written text, narrative or written dialogue, good style decrees that paragraph structure should reflect discourse and topic structure. Generally the beginning of a paragraph coincides with the beginning of a discourse segment, and a paragraph end marks the conclusion of a discourse segment, such as a topic or subtopic. However, spoken narrative and dialogue lack even these basic cues to discourse structure. We will look at two sets of features that have been shown to be correlated with discourse segment structure, specifically determining the location of discourse segment boundaries.

Acoustic Measures

The first group of markers of discourse structure are purely prosodic measures, measures of pitch, duration, or loudness rather than based on the lexical content of the utterance. Monologues in which a speaker gives directions from one location to another, called direction-giving tasks, have been used extensively in several studies of discourse structure. Such discourses are chosen because giving directios usually involves an overall goal with clear beginning and end, as well as clearly defined subgoals; this clear structure contrasts with the more muddled structure of a conversation between friends, for example. In one such study, [Davis and Hirschberg, 1988] examined direction-giving and found that pitch range was closely related to position in a discourse segment. Specifically, these researchers found that utterances initiating discourse segments had expanded pitch range, relative to utterances within the segment. These segment-internal utterances exhibited relatively compressed pitch range. These contrasts were then incorporated in a text-to-speech system, that demonstrated that expanded or contracted pitch range could lead subjects to different interpretations of the sequence of directions based on whether or not they perceived that a new segment had begun. A group of more systematic studies of intonation related to discourse structure were performed on a group of direction-giving monologues referred to as the Boston Directions Corpus. Studies of these monologues as reported in [Nakatani *et al.*, 1995] again noted an expansion in pitch range associated with segment beginnings and compression of pitch range within segments.

Finally in a single "dialogue" question-answer system study of user utterances in the ATIS (Air Travel Information System) corpus, a standard spoken language systems testbed, [Swerts and Ostendorf, 1995] explored the features of topic initiating utterances. This testbed is interesting in that, although it is a query system, user input is in spoken English, while output was presented in textual and tabular form. In addition to demonstrating the use of expanded pitch range in segment-initial position utterances in this new domain and new interaction style, [Swerts and Ostendorf, 1995] observed that there was an accompanying increase in utterance duration in segment-initial position.

These results have particular bearing on understanding the discourse role of corrections in spoken dialogue. Some researchers [Swerts and Ostendorf, 1995] have suggested that corrections are, in fact a form of segment-initial utterance. Specifically, by performing a correction the user is initiating a correction subdialogue or clarification subdialogue. This analysis would then argue that corrections should share many acoustic features with other discourse segment initial utterances. We will return to this question after our own analyses of corrections in the SpeechActs data examined in this thesis.

Acoustic and Textual Cues

In addition to the completely text-independent cues to discourse segment structure discussed above, a common textual feature identified as signaling discourse structure is the use of cue words or cue phrases. Cue words are words that perform a discourse function, possibly in addition to their innate semantic meaning. Common examples of cue phrases are "now", "OK", "first", etc. "Now" obviously means "at this time", but when used in its discourse cue sense as in "Now, what are we going to do about this correction problem?", it signals a shift in topic, that the following utterance

belongs to a different segment from the preceding statements. Likewise, "OK" can be used to signal a transition between two discourse segments. However, the problem of determining whether a potential cue word is being used in its cue sense or in its basic meaning remains, even when one has successfully recognized the presence of one the words in an utterance. [Hirschberg and Litman, 1993] analyzed occurrences of common cue words in both their cue and literal uses. They determined that the cue use of a word could often be identified based on simple acoustic measures. Specifically, the presence of an L* accent (a low pitch accent on a stressed syllable) on the possible cue word and a possible following silence signaled the discourse use of the word. Thus, in addition to general utterance acoustic characteristics, discourse segment structure can be determined based on the presence of certain word classes, cue phrases, with a specific intonation contour, raising the question of whether similar cues are available for identifying spoken corrections.

2.1.2 Inferring Specific Discourse Functions from Speech

Thus far, we have considered simply determining the beginnings and endings of discourse segments from speech features. Now we turn to inferring specific discourse relations. While the structural significance of corrections is somewhat unclear as to whether or not they should be considered to be segment-initial utterances, their functional, corrective, significance is both clear and of crucial importance. Clearly, some cue phrases, such as we discussed above, can carry specific discourse functions. "First" begins a narrative or list, "next" can continue a narrative sequence, and "however" can introduce a contrastive view. The use of specific phrases to identify given discourse relations has been studied extensively in text-based discourse analysis, such as Rhetorical Structure Theory (RST) by [Mann and Thompson, 1986] and augmented transition network discourse theories such as those by [Reichman, 1985]. However, as with paragraph segmentation, these cues are useful when available, but unfortunately are much less common in casual question-answer or direction-giving spoken interactions common in speech understanding domains, than in more formal written styles. For instance, in the SpeechActs data, only seven of the several hundred correction utterances contains a lexical cue such as "No" or "I meant." The question posed is thus how to identify a specific discourse function of a spoken utterance in the absence of explicit cue phrases.

The research of [Taylor, 1995], [Taylor *et al.*, 1996a], and the University of Edinburgh has explored this problem. This research uses the HCRC Map Task Corpus, a collections of interactions between pairs of speakers of whom one is designated the leader and the other the follower. Their task is, given two slightly different maps, for the leader to direct the follower to draw a map path on his or her own map which accurately reflects the one on the leader's map. This task was intended to elicit a lot of negotiation, conversation, and discussion of appropriate reference terms for objects on the maps. They define a set of game-theoretic discourse relations such as inform, query-yn, reply-yn, query-wh, reply-wh, clarify, etc. Each utterance is labeled with one of the functions.

Next they analyzed the pitch contour of each utterance according to the rise-

fall-continuation model [Taylor, 1995]. This analysis assigns a pair of numbers to each non-flat segment of the contour. The numbers correspond to the parameters for the parabolic shape which best fits each portion of the pitch contour. The series of contour values and the discourse relation label are used to trained a neural network to classify new utterance contours as one of the available relation classes.

This relation information is then incorporated into the recognition process in the following fashion. The incoming utterance is passed to the neural network classifier and is assigned a game move label. The recognition hypotheses are now restricted to those that can fulfill that game move. This shift of recognition focus based on recognized discourse relation improves overall recognition accuracy.

Now, none of the game moves directly corresponds to our notion of corrections. This mismatch is actually not surprising since explicit corrections of misrecognitions are much less frequent in human-human dialogue such as that captured in the Map Task data, than in human-computer interaction. In addition, [Taylor *et al.*, 1996b] observe that not all relations have clearly associated contours. Although there is no specific game move associated the phenomenon, there are many instances in the Map Task Corpus where one speaker acts to correct a misinterpretation on the part of the other conversant. It is likely that this type of correction would have many similarities to corrections of misrecognition errors in human-computer interaction. For instance, one would expect to observe contrastive use of pitch accent or insertion of a preceding silence associated with the word or phrase being corrected. The question of whether they take on the full array of clear speech characteristics, such as durational increase, requires further analysis, since many of the situations in which clear speech characteristics come to the fore involve interactions with conversants with some perceived deficit, as with young children, the hard-of-hearing, or computers; this is not the case for these human-human interactions. However, this research has strong connections to that reported in this thesis, by demonstrating that discourse relations can be identified through acoustic- prosodic information, and providing a data set for comparison of corrections in human-human and human-computer interaction.

Another segment of the ATIS-based discourse study of [Swerts and Ostendorf, 1995] directly addressed the issue of corrections. The researchers examined a set of correction utterances in the air travel information domain. They observed three sets of statistically significant contrasts between original requests and corrections. These contrasts were in the following measures: increased utterance duration (including changes in lexical content of the correction), decreased tempo or speaking rate, and a decreased interval of silence between the current and preceding utterance. The most effective detector was the number of content words in common between the current and immediately preceding user utterance. Some utterances showed insertion of a pitch accent, stress, within an intermediate phrase or on a function word, such as 'the' or 'a', that would ordinarily not receive such accent. However, none of these pitch accent measures or overall utterance pitch maximum or pitch range measures reached significance. This absence of pitch contrasts was unexpected on the analysis of corrections as fulfilling a segment-initial discourse role.

2.2 Self-repairs

While relatively little research has been done on recognizing corrections of errors made by other conversational participants, a substantial amount of research has looked at recognizing self-repairs. Self-repairs are corrections within a single utterance, such as false starts, disfluencies, and the like. Consider the following utterance, "When does the plane for Austin leave on Mon - Tuesday?" This utterance contains a self-repair where the speaker changed the data of departure from Monday to Tuesday, after uttering the first syllable of the incorrect date. The goal of recognizing a self-repair is to pass to the natural language understanding module a string which corresponds to the corrected form of the utterance with all trace of the corrected, erroneous segment removed. In the literature, the corrected segment is referred to as the reparandum, and the new segment is called the repair. To accomplish this main task, it is necessary to identify the two regions constituting the reparandum and the repair. A variety of approaches to this problem have been proposed, and as in discourse segmentation and function inference, they can be divided into two groups based on whether or not the techniques rely on textual information.

2.2.1 Recognizing Self-Repairs with Text and Acoustic Information

Approaches reported by [Heeman and Allen, 1994] and [Nakatani and Hirschberg, 1994] make use of lexical and acoustic features to identify the locus of self-repairs in utterances. The lexical component of these methods involves finding matched sequences in the reparandum and repair, as in "Take the oranges to Albany - to Erie" where the word to starts both reparandum and repair. The presence of filled pauses, e.g. "umm" and "uh", and unfilled pauses is also found to be a useful cue to the initiation of a self-repair. The most important measures for these approaches were the presence of word fragments. When this measure was excluded measures of presence and length of pause proved to be the best signals to the presence of self-repairs. A lexical match within a three word window to the right and position within the utterance were also useful. [Nakatani and Hirschberg, 1994] found, in addition, that there were significant increases in both pitch and amplitude between the last stressed syllable of the reparandum and the first stressed syllable of the repair. This decrease in amplitude and cliticization or deaccenting of the word preceding the interrupting correction segment played a role in some classification in the Classification and Regression Tree (CART) used to find the boundary between reparandum and repair.

2.2.2 Recognizing Self-Repairs with Acoustic Information Alone

While they recognize that the best results for identifying self-repairs are achieved when textual information is used, [Shriberg *et al.*, 1997] describe an approach that uses only acoustic cues. They argue for this approach based on the fact that in the context of speech recognition an accurate text transcription of the full utterance is not necessarily available, particularly in the cases of utterances involving self-repairs

where the presence of disfluencies can lead to misrecognitions. Their approach considers each inter-word position to be a possible start of a self-repair. They then use information about the duration, pitch, amplitude, and pauses in preceding and following words to determine which positions are most likely to be associated with self-repairs. The most important features for identifying self-repairs or disfluencies were duration, distance from pause, and for certain disfluency classes, pitch, which played a more significant role than amplitude, as measured by signal-to-noise ratio (SNR), or gender.

These studies of self-repairs show that a form of correction, where one corrects oneself mid-utterance, can be recognized best based on repeated phrases within an utterance, but can still be identified with some success based on acoustic features such as duration and amplitude. These approaches suggest several acoustic measures to be used in identifying corrections of errors made by either conversant. However, corrections are likely to prove more difficult to recognize both because repetition of words between utterances, even in the absence of correction, is more likely than within utterances and because this direct comparison of original and correction may not be possible.

2.3 Corrections

Recent research by [Oviatt *et al.*, 1996] has focussed on characterizing the phonetic, acoustic, and prosodic changes that take place in spoken corrections. The data in [Oviatt *et al.*, 1996]'s study was collected in a simulation study where randomly generated errors were presented to a subject who was using a speech interface to a form-based interface where recognized input and recognition failures were signaled visually on a WACOM graphical template. Thus there was a voice-in/tabular-visual-output system. While [Oviatt *et al.*, 1996] do not make any attempt to automatically recognize these spoken corrections, they carefully analyze and characterize the spoken corrections collected in the experiments. First they identify a cluster of phonetic changes between otherwise lexically identical original inputs and repeat corrections. These changes, such as shifts from flapped to released t's reflect a change from a more conversational to a more clear and precise speaking style in 10% of correction utterances. They also found significant increases in total utterance duration, speech duration, number and length of pauses, and decrease s in pitch minimum among male speakers. However, no changes in amplitude were observed. These increases in utterance and pause duration fit into the same contrast of a shift from a more conversational to a more careful, clear speech style between original and correction as the phonological changes above.

The research in this thesis is a natural and more computational extension of the work by [Oviatt *et al.*, 1996] in which the author participated. The prior work provided the insight of casting corrections as instances of hyperarticulation, that can be characterized by a specific set of acoustic adaptations. That research also developed an analysis methodology for comparing different utterance classes, in this case original and repeat, systematically for a variety of acoustic-prosodic measures. Many similar

measures, along with extensions to speaking rate and refinements, such as per-subject normalization, play a primary role in the analysis in this thesis. Discussions of the best ways to exploit the contrasts between corrective and non-corrective utterances led to the decision to use them to build a correction classifier and adapt automatic speech recognition.

2.3.1 Speaking Styles

Since the above work on corrections has raised the issue of treating corrections as shifts from one speaking style to another, let us consider briefly the work of [Ostendorf *et al.*, 1996] in a summer workshop held at Johns Hopkins University in 1996. This workshop focussed on improving speech recognition rates by incorporating a speaking mode variable into the speech recognition model. This mode variable was intended to capture some of the systematic contrasts in phonological features that accompany differences between read and conversational speech. These differences are often blamed for the relatively poor accuracy rates on recognition of casual conversational speech as found in the Switchboard or "call-home" corpora, where people were recorded making free-form telephone calls to their family and friends, reaching about 40% word error rate, versus the 5-10% word error rates achieved on read Wall Street Journal text. For the read versus conversational contrast the designers built a decision tree classifier that could discriminate between these classes at an error rate of 27-34%. The classification was based most heavily on durational and speaking rate measures. In addition, presence of a pause of length greater than 50 milliseconds and loudness measures, in terms of signal-to-noise ratio (SNR), also played a role. Pitch measures did not provide any significant information to the classification process. They also found that they could significantly improve accuracy on the more difficult conversational corpus data by collecting a set of common phonological changes and using these to modify the recognition model for words in utterances classed as conversational by the decision tree or other method.

This work has some clear parallels with that discussed in this thesis, in that it tries to use acoustic measures and classifiers to recognize certain speaking styles and seeks to understand the impact of the speaking style on the recognition process. However, we will see that the work in this thesis looks at identifying a *specific set* of features that identify an important discourse role, marking corrections, which has proven difficult even with perfect transcriptions and has proved opaque in the absence of this lexical information. In addition, at a finer granularity, we find different features to play a role in signaling corrections from those involved in the read-conversational speaking style contrasts. We seek to relate these contrasts to those for other discourse functions in the hope of better understanding the role of corrections in discourse structure as a whole.

Chapter 3

Data Collection: System, Subjects, and Overall Analysis

The data used in the experiments and analyses throughout this thesis was drawn from recordings of user interactions with the Sun Microsystems SpeechActs system during a limited field trial. This chapter will provide an introduction to the SpeechActs system itself, a general description of the data obtained in the field trial, a discussion of the identification, coding, and analysis of errors encountered by system users, and a description of the criteria by which the utterance pairs used in the acoustic analysis and recognition experiments were selected.

3.1 SpeechActs Description

SpeechActs is a prototype spoken language system designed by Sun Microsystems Laboratories Speech group (1993-1995), where the author spent two summers as a student intern. SpeechActs provides a voice-only interface to a variety of common desktop and information feed services. Specifically, it included e-mail reading and sending, access to one's own and other browsable calendars, weather information, stock quotes, time zone and currency conversions, and a notification system. An example interaction is shown below:

SpeechActs: You have 14 new messages and
47 old messages. The first new
message is from Eric Baatz
regarding "report completed."
User: Let me hear it.
SpeechActs: "The first draft is ready for
your comments. Eric."
User: Reply and include the current message.
SpeechActs: Preparing message to Eric Baatz.
Begin recording after the tone.
When finished, pause several seconds.
User: Eric, I'll get those to you
this afternoon.
SpeechActs: Say cancel, send, or review.

User: Send.
SpeechActs: Message sent. What now?
User: Next message.
SpeechActs: New message two from Stuart Adams, subject "Can we meet at 3:00 today?"
User: Switch to calendar... [1]

Two significant features distinguish SpeechActs from other spoken language systems. First, since it was designed primarily for use over the telephone to provide ubiquitous access, it is a voice-only system. With the exception of password entry and escape sequences which use DTMF input, all user input is spoken, and all output is through synthesized speech; there are no visual displays for feedback. Speech recognition is performed by BBN's Hark speaker-independent continuous speech recognizer, and synthesis is performed by Centigram's TruVoice text-to-speech system. Locally developed natural language processing and interpretation engines feed the speech recognition results to the appropriate speech application interfaces for each backend system. The restriction to a single, audio modality requires that the user give all necessary information to the system through speech, and allows our analysis to focus exclusively on those speech cues of lexical, phonetic, and acoustic-prosodic form which the spoken modality provides.

Secondly, SpeechActs was designed to provide a "conversational" interface. A conversational interface can best be understood by what it is not. It is not a fixed command language, it is not a form-based input structure, and it does not have rigid vocabulary or syntax. Instead, a conversational interfaces hopes to provide both ease of use for novice users and efficiency for more experienced users by allowing them to use language which comes naturally for each individual. In addition, it is easy to combine commands or criteria for requests into a single command for more confident and experienced users (e.g. read the third urgent message) or to simply step through the information with a sequence of simple commands for novice users (e.g. "Go to urgent messages","Next","Next","Next"). All new users are provided with a wallet-sized information card with examples of common commands for each application, but, as we will demonstrate later in this chapter, users each rapidly develop their own distinct style and vocabulary.

3.2 Data Collection and Coding

Now that we have provided a general overview of the SpeechActs system, let us turn to a more detailed description of the data collection process. As discussed above, SpeechActs was deployed for a limited field trial over an analog telephone connection, so that it could be accessed from home, office, hotel, or even a busy, noisy airport terminal. All interactions were recorded automatically during the course of the conversation. All speech, both user input and system synthesized responses were

digitized and stored at 8kHz sampling rate in 8-bit mu-law encoding on a single channel, compatible with native system hardware and the limitations of analog telephone lines. In addition to the stored audio, speech recognizer results, natural language analysis results, and the text of all system responses was recorded and time stamped.

Next, all user utterances were textually transcribed by a paid transcriber. Each transcription of user input was paired with the speech recognizer output for that utterance. Each of these pairs was assigned one of four accuracy codes:

- Correct: Recognition and Action Correct

  User Said: Read message one

  System Heard: Read message one

- Error minor: Recognition not verbatim; action correct

  User Said: Go to the next message

  System Heard: Go to UH next message

- Misrecognition: Recognition not verbatim; action incorrect

  User Said: Next

  System Heard: Fax

- Rejection: No recognition; no action

  User Said: Read message one

  System Heard: nothing

The use of the "Correct" code should be evident. The "error minor" code assignments generally resulted from a misrecognition of a non-content word (e.g. wrong tense of an auxiliary verb, incorrect article, insertion of "um" or "uh") for which the robust parsing of the natural language component could compensate. The "misrecognition" and "rejection" codes were assigned in those cases where a user could identify a failure in the interaction. Utterances coded either as Misrecognition or Rejection could also receive an additional tag, OOV. This tag indicates that either words not in the recognizer's vocabulary or constructions not in the systems's grammar were used in the utterances. For simplicity, however, we refer to all these cases as OOV. Two examples appear below:

- Unknown Word: Rejection

  User Said: Abracadabracadabra

  System Heard: nothing

- Unknown Construction:Misrecognition

  User Said: Go to message five eight six

  System Heard: Go to message fifty six

  Grammar knows: Go to message five hundred eighty six

In total, there were 7529 recorded user utterances from the field trial. Of these. 4865 were correctly recognized by the speech recognition pass, and 702 contained minor recognition errors, but still resulted in the desired action. There were 1961 complete recognition failures: 1250 of which were rejection errors and 706 of which were substitution misrecognition errors. The remaining errors were due to system crashes or parsing errors. In other words, almost two-thirds of recognition failures were rejections, about twice the number of misrecognitions. [2] Overall, this results in a 25% error rate.

We also observe, like [Shriberg *et al.*, 1992], that there is a higher probability of a recognition error following an error than following a correct recognition. Specifically, the probability of an error after a correct recognition is approximately 18% whereas after a recognition failure it rises to 44%, more than 2.75 times as likely. This contrast is evident in the presence of, often lengthy, error spirals in which multiple errors follow a single initiating error. This contrast in recognition accuracy between original and correction utterances motivates the contrastive analysis which follows and efforts to characterize the changes which mark corrections.

3.3 Longitudinal Change, OOV errors, and novice-expert contrasts

The subjects participating in the field trial included fourteen individuals drawn from the Sun Microsystems sales, marketing and technical staff with no previous experience with spoken language systems, four members of the SpeechActs development staff, and a group of one-time guest users who called in to try out the system. There were three female and fifteen male regular system users. The users engaged in at least ten phone conversations with the system. The distribution of users allows us to examine the development of novice users' interaction style, in terms of vocabulary choice and number of out-of-vocabulary (OOV) utterances. In addition, we can contrast the different recognition accuracy rates and vocabulary distributions of expert and novice users.

3.4 Vocabulary Changes

We have observed that interactions with the SpeechActs system resulted in a 25% error rate. We would like to understand how these errors are distributed across users. Since we have as subjects both expert developers and novice users just learning how to use the system, we can compare error rates for these two groups. How important are the errors and error rates we have observed? If all of the errors are produced by very early novice users, one might choose to ignore the issue of errors as only a passing problem. However, if errors persist, it is particularly important to understand how

[2] Curiously, this ratio of rejection errors to misrecognition errors is reversed from that most often reported in spoken language systems. The relatively high rate of rejection errors may be attributed to the noisy telephone environments in which this system was most often used.

errors arise and how to handle corrections. We will find that novice users improve their ability to interact with the system over time, but still encounter many more errors than expert users. We will explore the factors that contribute to this improvement by focusing on the most glaring, if expected, difference between new and expert users: the use of words and constructions that are outside the system's vocabulary. We will look at the change, fortunately decreasing, in OOV utterances by novice users over time. We will examine how the reductions in unknown utterances are expressed in terms of the working vocabulary size of these users.

novice users show significant decreases in vocabulary size and rate of introduction of new words. Developers and one-time users demonstrate no such changes. Finally we observe that in spite of the small final vocabulary sizes reached by users, fewer than 50% of words are shared between any two users. While the vocabulary of any one user is quite small, a much larger vocabulary is needed cope with variation between users.

3.4.1 Error and OOV Rates

Let us begin with the question of which users are making the errors that give rise to an overall 25% error rate. We compute overall average error rates for each subject, novice and expert. Figure 3-1 displays the distribution of overall average error rates for all subjects, with novice users and developers plotted separately. Next we compute the overall rate of out-of-vocabulary utterances for each subject in the two groups, shown in Figure 3-2 A comparison of novice users with system developers indicates a significantly higher rate of overall recognition (24.86% vs 10.75%) and OOV (7.39% vs 0.76%) errors for novices than for system developers.

The next important question to address is whether these error rates, especially the higher novice user error rates, change over time, and if so, how and how much. To track these longitudinal changes, or changes over time, we recompute the error and OOV rates from above in terms of number of errors per hundred utterances for the first, second, and third set of one hundred utterances, and so on. For each time point, or group of 100 utterances, we present a box-and-whisker plot showing the range of error rates for all novice users (Figure 3-3), all expert users (Figure 3-4), and all single shot "guest users" (Figure 3-5).

We can see that neither the expert users nor the single shot users show any particular change in error rate over time. However, novices show a distinct decrease in errors from the first hundred utterances to the second hundred to a relatively stable and lower error rate. We can quantify this contrast by comparing number of errors in the first hundred utterances to the average number of errors per hundred utterances for the later interactions. (Figure 3-6) This contrast is a significant decrease by t-test, one-tailed. ($t = 2.07$, $df = 22$, $p < 0.05$) , showing that novice users make fewer errors over time, but still at a much higher rate than expert users.

This observation comes as no surprise; however, we would like to know which features of novice vs. developer interaction account for this contrast. Specifically, to what degree do out-of-vocabulary utterances or speech acoustics differentially affect the error rates of these two subject groups? Can all contrasts be related to limited

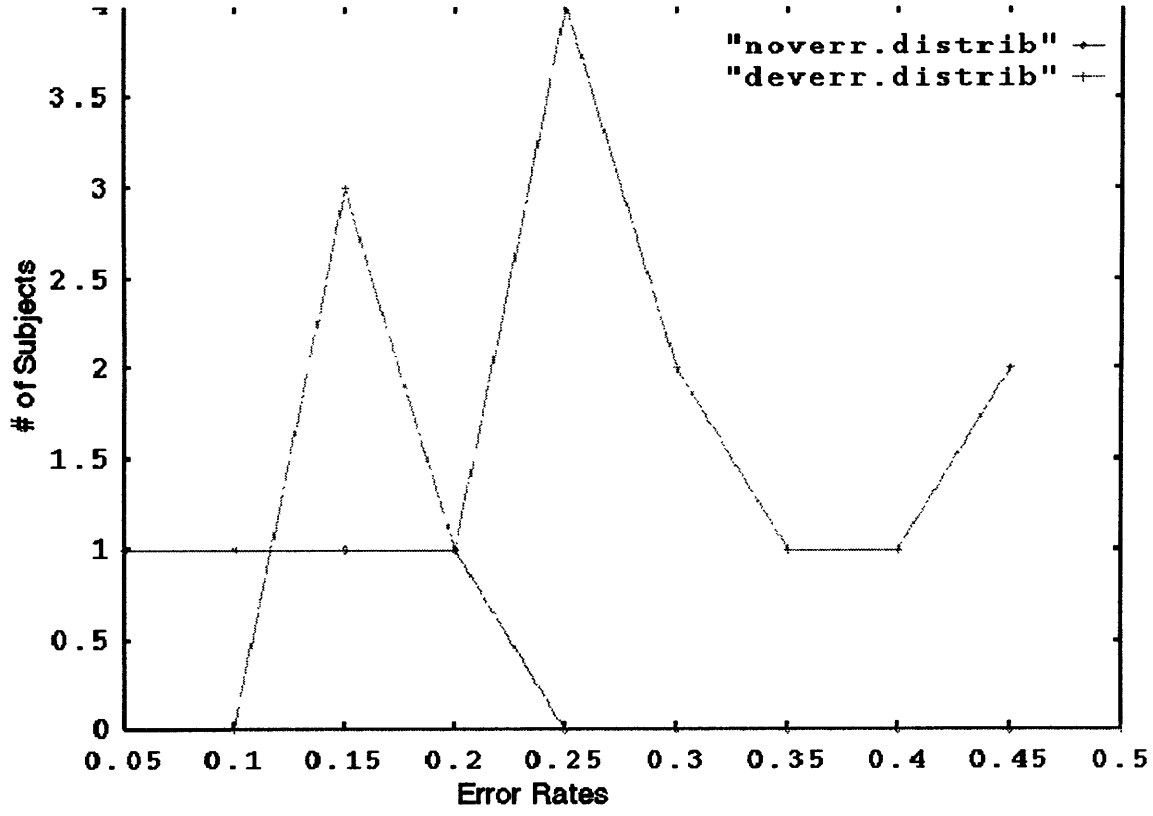


Figure 3-1: Distribution of average error rates for novice (light grey line) and expert users (dark grey line): Developers produce recognition error rates between 3% and 16% (10.75% on average), while novice users experience much higher error rates, between 12% and 43% (24.86% on average).

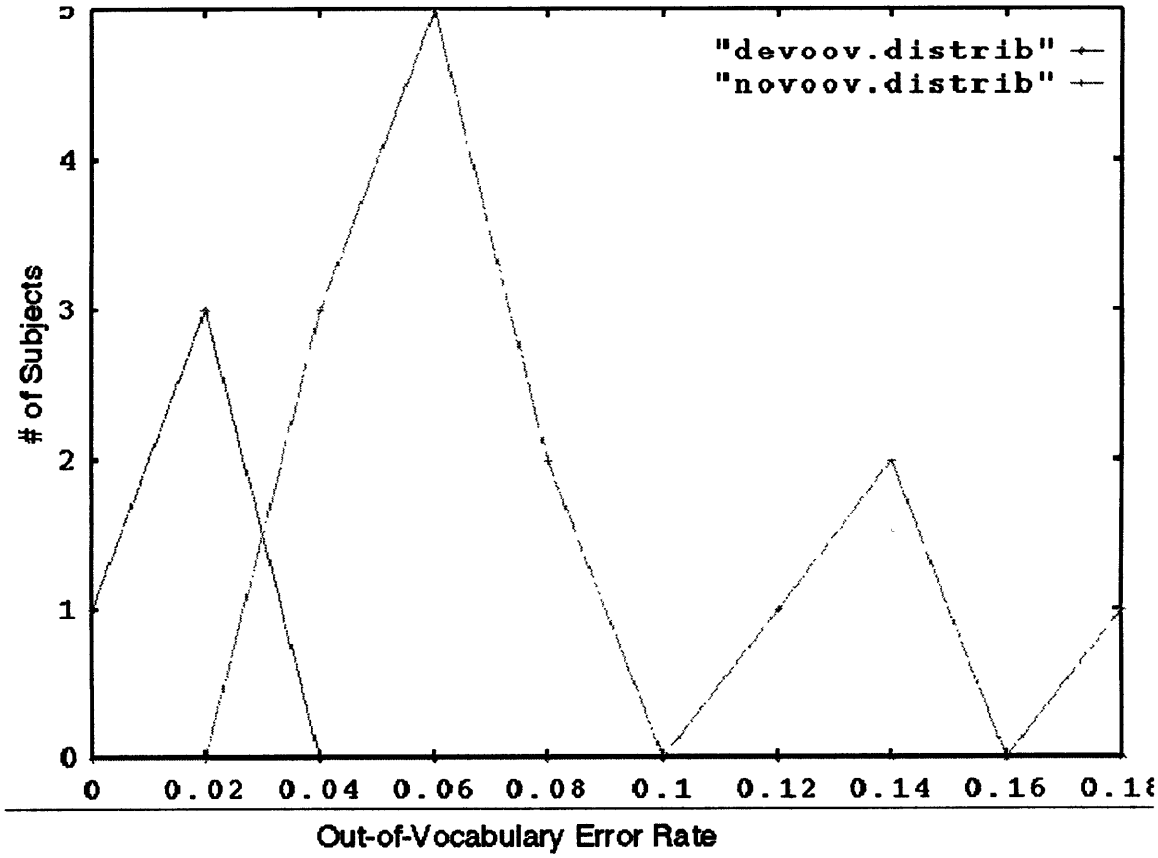


Figure 3-2: Distribution of average OOV rates for novice (light grey line) and expert users (dark grey line): Expert users rarely produce out-of-vocabulary utterances, accounting for between 0 and 2% of utterances (0.76% on average). Novice users in contrast use utterances not understood by the system in 3 - 18% of utterances (7.39% on average.

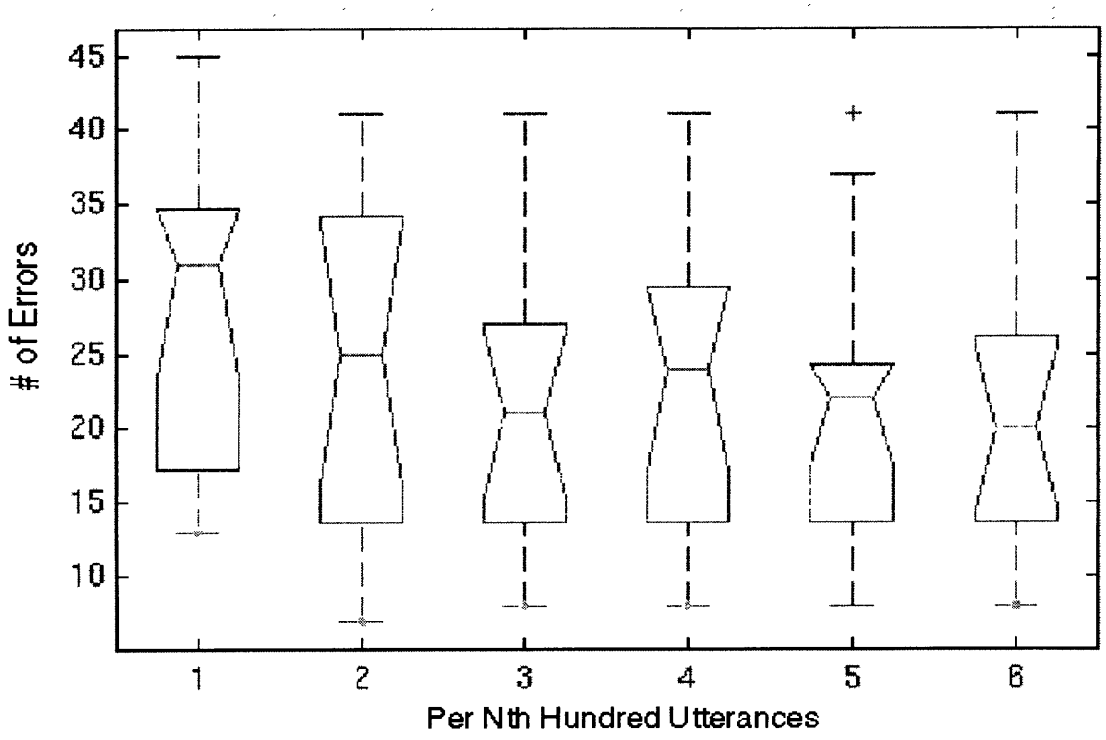


Figure 3-3: Novice Users: Errors per hundred utterances: Over time, novice users encounter fewer recognizer errors with most of the improvement taking place over the first 300 inputs to the system.

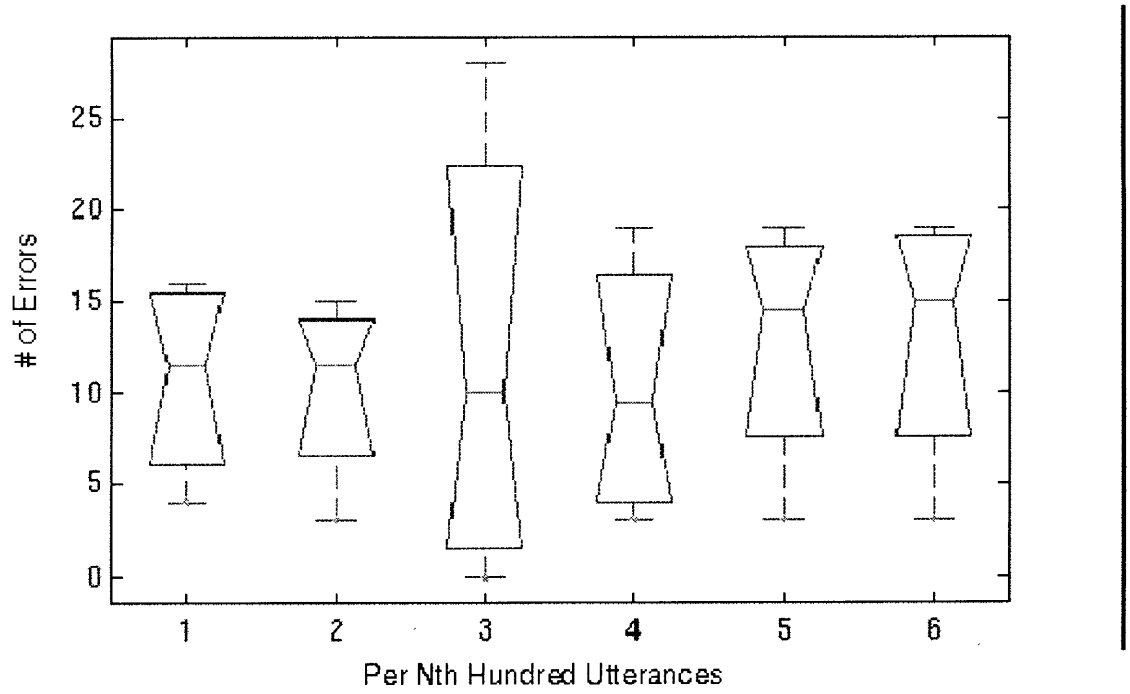


Figure 3-4: Expert Users: Errors per hundred utterances: Over time, expert users remain relatively constant in the number of recognizer errors they incur.

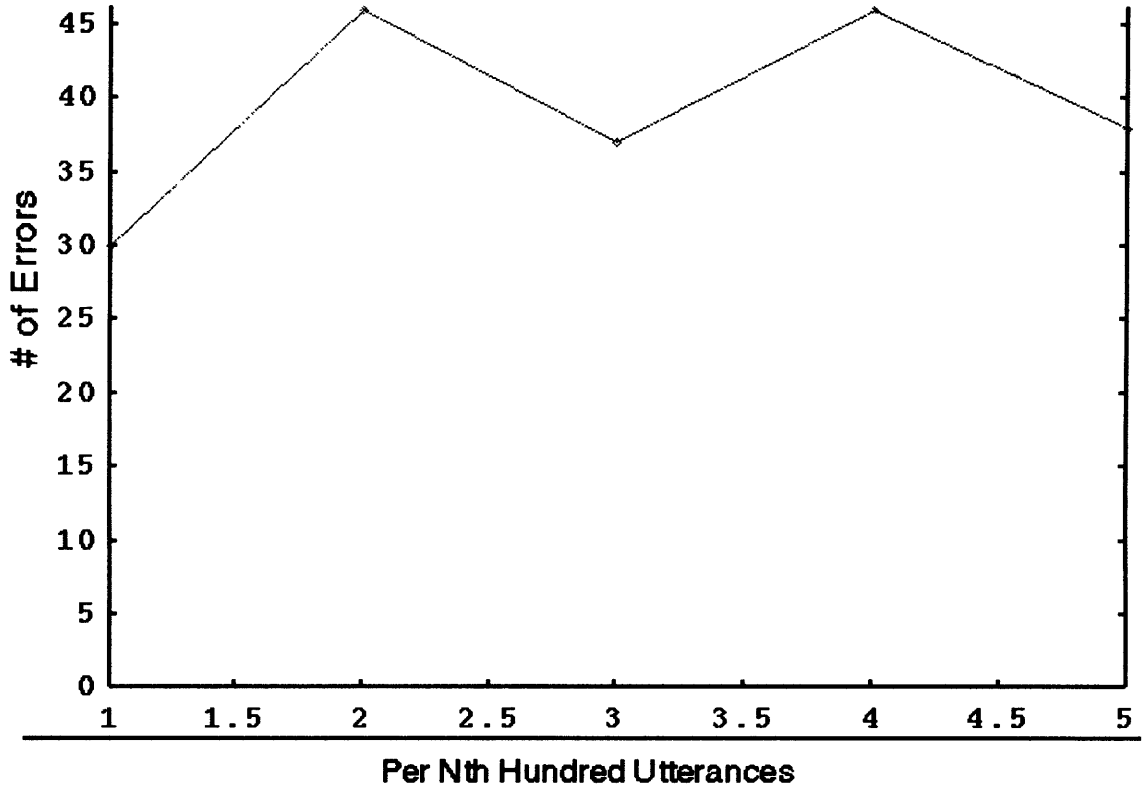


Figure 3-5: Single Shot Users: Errors per hundred utterances: Unsurprisingly, single shot guest users of the system also show no improvement in recognition accuracy.

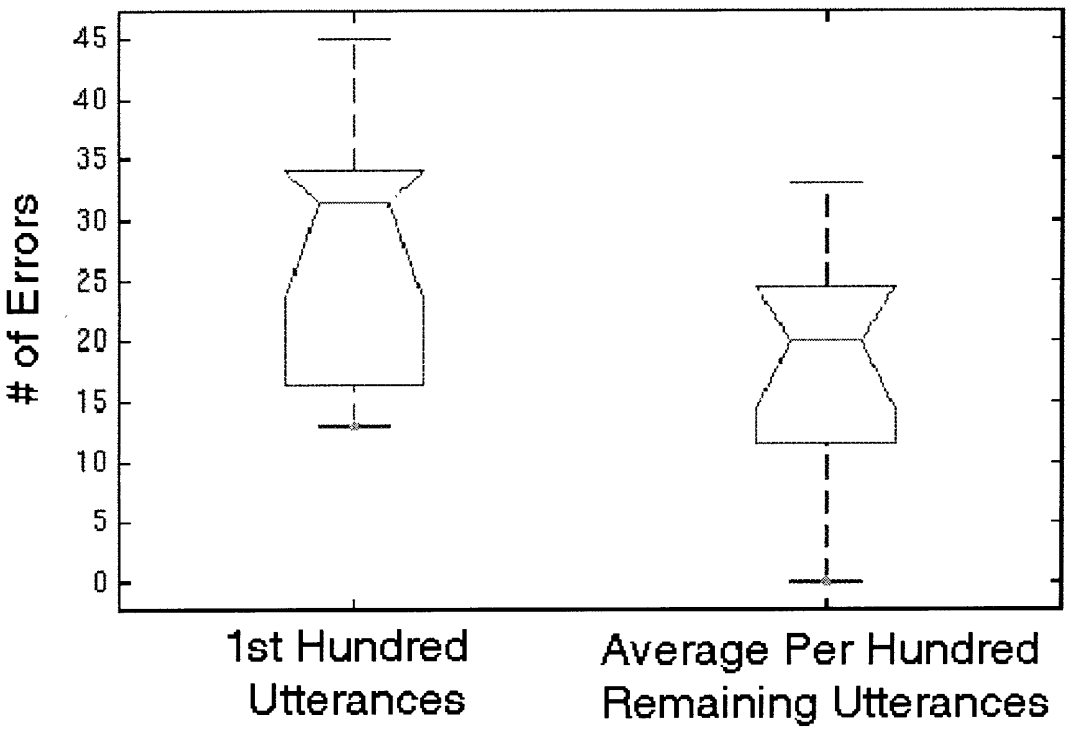


Figure 3-6: Novice Users: Errors per hundred utterances: First hundred versus Average: More concisely, we contrast the number of recognizer errors encountered by novice users in their first hundred interactions with the average number of per hundred utterances in later interactions. There is a significant decrease in error rate over time for these users.

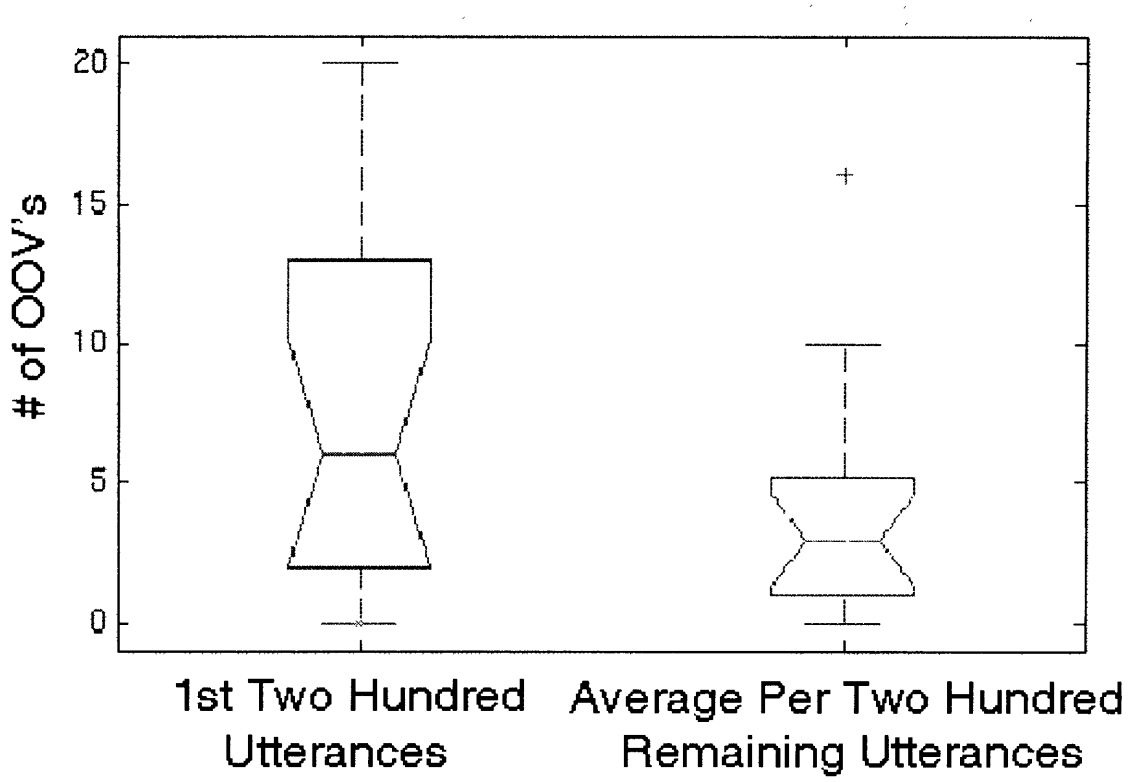


Figure 3-7: Decrease in out-of-vocabulary utterances over time: Again, for clarity we contrast the number of out-of-vocabulary errors by novice users in their first 200 utteranceds with the average number of such errors in later interactions. Here too we find a significant decrease in the number of illegal utterances by novice users as they gain experience with the system.

knowledge of the system's vocabulary? Experts, naturally, exhibit very few instances of out-of-vocabulary utterances. Here we consider the change in rate of OOV's in novice user utterances over time and contrast it with that of the guest user class. There is a significant decrease in OOV's over time for longer term users, in contrast with an almost constant OOV rate for single-shot users. Specifically there is a significant (T-test, two-tailed, $t = 2.3$, $df = 32$, $p < 0.05$) decrease the number of OOVs between the first 200 utterances and all subsequent interactions. Figure 3-7 demonstrates this drop in number of out-of-vocabulary utterances. ANOVA shows a significant effect of number of interactions. ($F(1,32) = 5.171$, $p < 0.03$) This is clearly a desirable trend, indicating the new users' increasing familiarity with the limited vocabulary understood by the system.

However, by comparing error rates in the first hundred utterances to the average of subsequent hundred utterance sets, we see that when these figures are computed without the errors contributed by OOV-related errors, the decrease in error rates with time is not significant. ($F(1,22)= 0.7246$) (Figure 3-8) The decrease in OOV errors is thus the primary contributor to the perceived improvement in recognition rate over time. In addition, even with all OOV errors removed the error rates of novices are

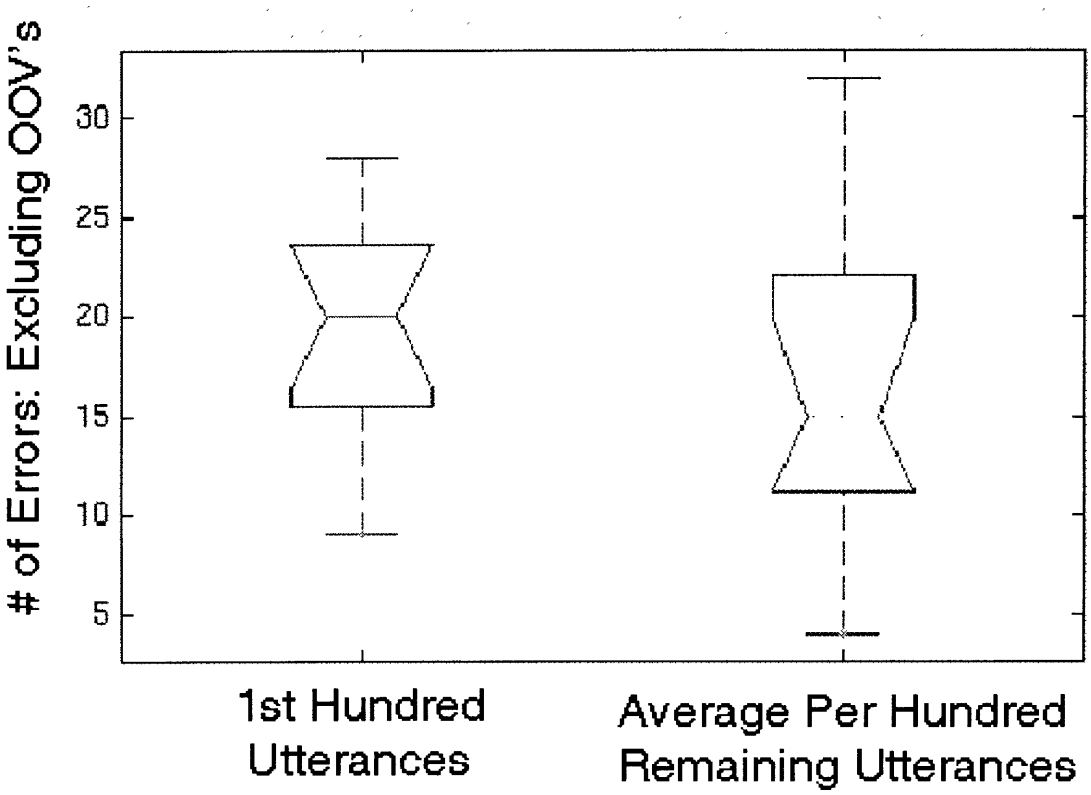


Figure 3-8: Novice Users: Error Rates for non-OOV errors: Here we exclude errors due to out-of-vocabulary inputs to compare novice error rates over time. While the average number of errors does decrease between the first hundred utterances and the average for later interactions, this decrease is not significant. Most of novice user improvement in recognition is due to increased knowledge of system vocabulary.

still much higher than those of expert users (18.25% versus 10.25%), indicating that expert use of a spoken language system requires more than just the knowledge of the utterances understood by the system. This knowledge is acquired fairly rapidly as we see by the drop in OOV rates, but the knowledge of proper speaking style is more difficult. (Figure 3-9)

3.4.2 Vocabulary Size and Rate of New Word Introduction

The next question to address is how to account for this decrease in OOVs. Does the user simply replace unknown word instances with known words? Does the user's working vocabulary increase, decrease or stay the same? Here we will use two measures to try to clarify the process of OOV reduction: number of words in working vocabulary (defined as number of discrete words per hundred words spoken) and rate of introduction of new words into the working vocabulary (again in words per hundred). Unsurprisingly, the rate of new word introduction undergoes a significant decrease over time - for all except the guest user category - and, like OOVs, drops

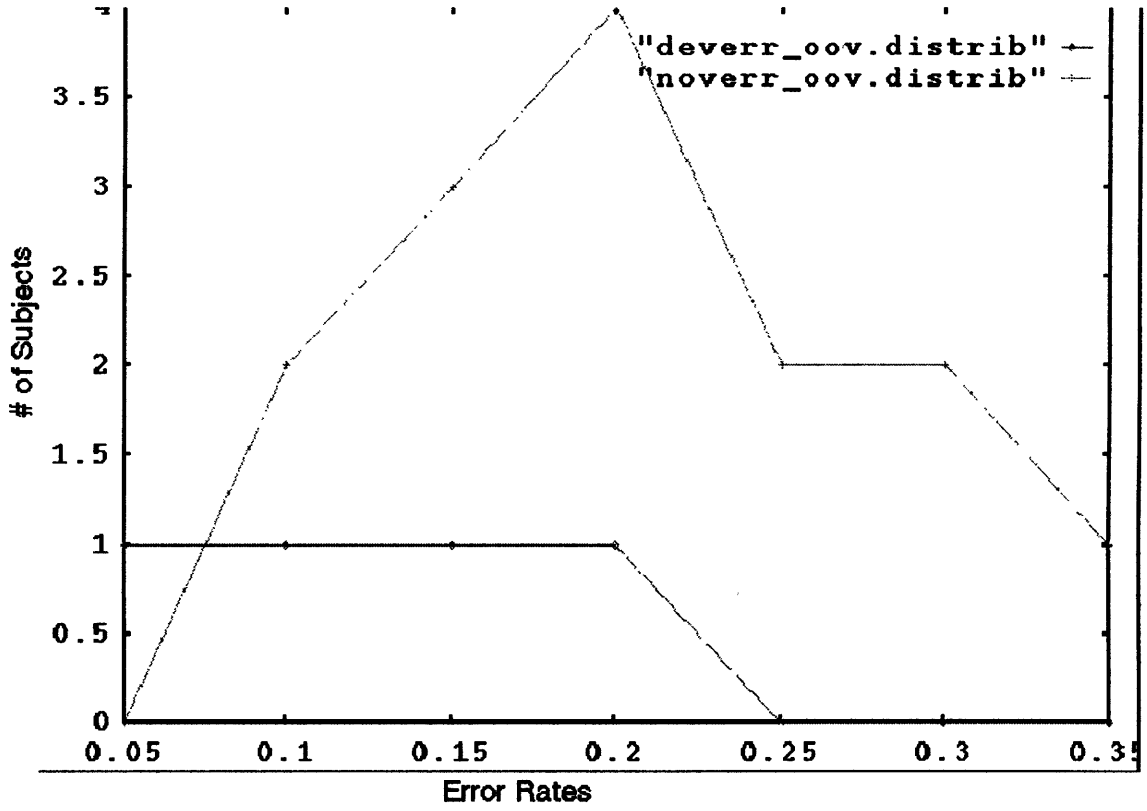


Figure 3-9: Distribution of Error Rates Excluding OOVs: novice users (light grey line) versus expert users (dark grey line): When errors due to out-of-vocabulary utterances are ignored, almost one-half of novice users achieve similar error rates to experts. However, the average error rate for developers at 10.25% is still much lower than that for novice users at 18.25%.

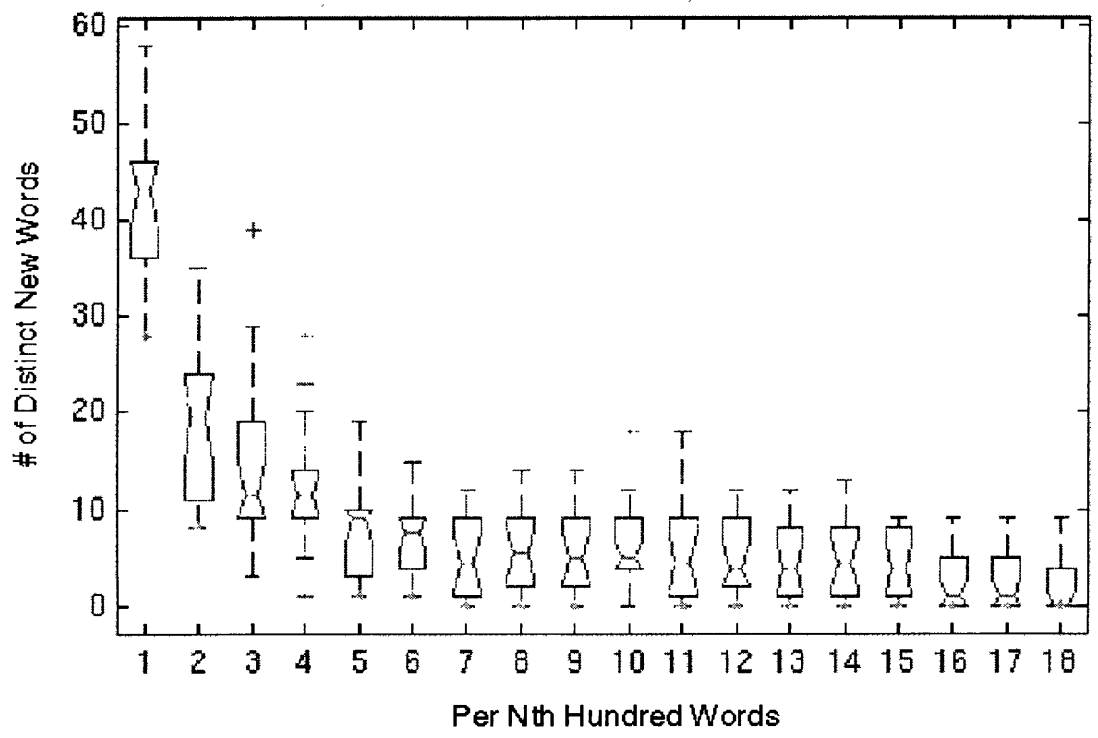


Figure 3-10: Distinct new words per hundred words: Initially, novice users introduce large numbers of new words into their vocabulary. However, this rate levels off over time, particularly after the first 600 words. This decrease in introduction of new words is a significant trend.

dramatically after the first 200-300 words. This trend is displayed in Figure 3-10.

Analysis of variance of number of new words to point in time is highly significant. ($F(17,306) = 59.27$, $p < 0.001$)

The trend for the working vocabulary is quite interesting and somewhat unexpected. Again, paralleling the decrease in word introduction, there is a significant decrease in vocabulary size over time. Specifically, there is a significant decrease in the number of unique words per hundred between the first 200-300 words and all later interactions. ($F(1,18) = 8.738$, , $p < 0.01$) Figure 3-11, Figure 3-12Curiously, the novice users each seem to converge on a fairly small vocabulary of 30-40 unique words per hundred. Specifically, novice users, after working with the system for an extended period of time, converge on a working vocabulary of an average of 35 distinct words per hundred, in a strong contrast to the 50 distinct words per hundred of the developer set.

From these analyses, we can see that the decrease in out-of-vocabulary utterances arises from a narrowing of the users' working vocabulary to a fairly small set of words in which the user has high confidence.

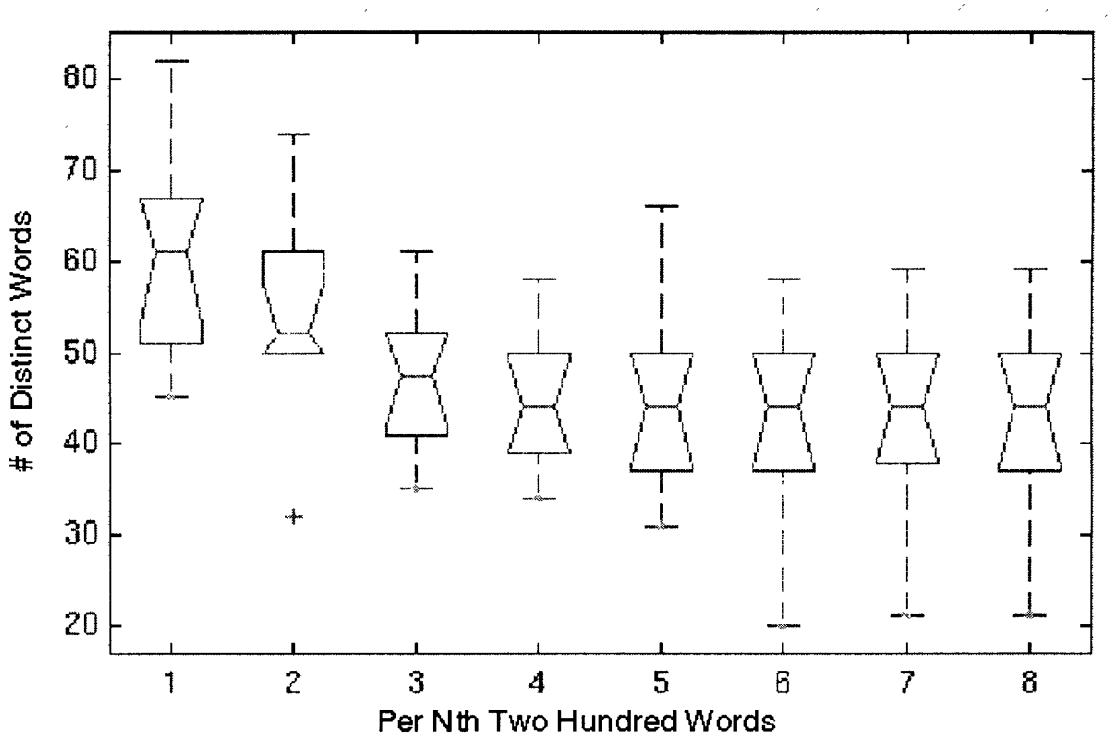


Figure 3-11: Distinct words per 200: allsubjects: Over time, not only does the number of new words decrease, but the actual size of the user's working vocabulary decreases.

3.4.3 Vocabulary Overlap

What ramifications does this use of a small working vocabulary have for conversational speech user interface design? Is it simply irrelevant since only a small set of words is needed by any user? An analysis of cross-user vocabulary will help to answer these questions. Here we tabulated the percentage of words shared between any pair of users and the percentage of a user's vocabulary which overlaps with any other's. We see that, for any pair of users, between 18 - 57% of vocabulary is held in common, with an average of 21% of the union of the two vocabularies falling in the intersection. Table 3-13 This translates to each user sharing approximately 50% of their words with any other given user. Table 3-14 This relatively small proportion of overlap between users attests to the value of the conversational interface. While the users individually do not have large vocabularies, the choice of words across users is highly varied. This supports the notion of a flexible vocabulary that allows users to gravitate toward lexical usages which come naturally, and supports wide cross-user utility. It is difficult to determine the exact criteria by which users select their final vocabulary. It was suggested that the users might be choosing those words that are not misrecognized by the system; in other words, the users are being trained to a certain set of usages by their success with the system. However, in examining the data, we observe that users persist in employing words and expressions that are often misrecognized by the system. In fact, most of the words in a user's

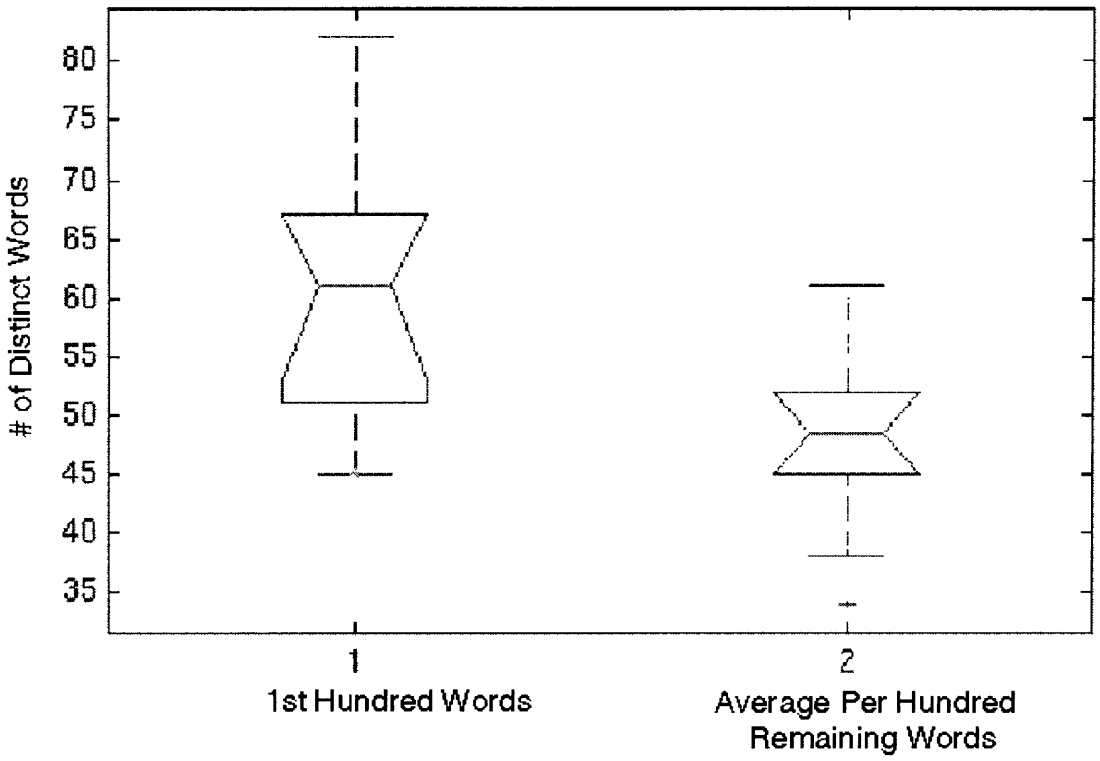


Figure 3-12: Novice Users: Distinct words per 200: first 200 versus remainder: More concisely, we see that the size of a novice user's working vocabulary, the number of unique words per hundred uttered, decreases significantly between the earliest interactions and later speech.

	Subjects								
Subjects	1	2	3	4	5	6	7	8	9
1	1.00	0.30	0.44	0.48	0.41	0.48	0.30	0.37	0.41
2	0.21	1.00	0.53	0.34	0.26	0.34	0.34	0.42	0.37
3	0.19	0.32	1.00	0.22	0.24	0.27	0.21	0.32	0.24
4	0.33	0.33	0.36	1.00	0.26	0.36	0.36	0.28	0.33
5	0.42	0.38	0.58	0.38	1.00	0.31	0.31	0.35	0.31
6	0.41	0.41	0.53	0.44	0.25	1.00	0.38	0.38	0.44
7	0.33	0.54	0.54	0.58	0.33	0.50	1.00	0.33	0.46
8	0.33	0.53	0.67	0.37	0.30	0.40	0.27	1.00	0.40
9	0.37	0.47	0.50	0.43	0.27	0.47	0.37	0.40	1.00

Figure 3-13: Vocabulary Overlap Ratio: $\frac{\|\bigcap(Vocab(subj_x),Vocab(subj_y))\|}{\|Vocab(subj_x)\|}$

	Subjects								
Subjects	1	2	3	4	5	6	7	8	9
1	1.00	0.14	0.15	0.25	0.26	0.28	0.19	0.21	0.24
2	0.14	1.00	0.25	0.20	0.19	0.23	0.27	0.31	0.26
3	0.15	0.25	1.00	0.16	0.20	0.22	0.18	0.27	0.19
4	0.25	0.20	0.16	1.00	0.18	0.25	0.29	0.19	0.23
5	0.26	0.19	0.20	0.18	1.00	0.16	0.19	0.19	0.17
6	0.28	0.23	0.22	0.25	0.16	1.00	0.27	0.24	0.29
7	0.19	0.27	0.18	0.29	0.19	0.27	1.00	0.17	0.26
8	0.21	0.31	0.27	0.19	0.19	0.24	0.17	1.00	0.25
9	0.24	0.26	0.19	0.23	0.17	0.29	0.26	0.25	1.00

Figure 3-14: Vocabulary Overlap Ratio 2: $\frac{\|\bigcap(Vocab(subj_x),Vocab(subj_y))\|}{\|\bigcup(Vocab(subj_x),Vocab(subj_y))\|}$

working set have been misrecognized by the system at some time in their interactions. Specifically, for novice users, we find that between 80 and 94% of their working vocabulary was misrecognized during their interactions. For expert users, with more varied vocabulary and lower misrecognition rates, about 57% of a speaker's working vocabulary on average had been misrecognized by the system. Clearly, the process of vocabulary convergence is much more complex than a simple trigger process, where a recognition failure causes the user to discard the vocabulary item.

3.5 Pair Data Selection

The remainder of this thesis focusses on characterizing and recognizing correction utterances in contrast other inputs. In order to provide a minimal clear contrast, following Oviatt et al, we consider lexically matched original input-repeat correction pairs. Specifically, we select pairs of user utterances which precede and follow a system recognition failure. The first user utterance, the original input, is a first attempt to

input a command or piece of information. The second utterance in the pair, the repeat correction, is the utterance immediately following the system response indicating a recognition error, either a rejection or an inappropriate response. which attempts to reinput the same command as the original. We choose only lexically matched pairs, user utterance pairs with the same word sequence in both original input and repeat correction. This constraint allows us to limit variation in acoustic measures due only to lexical or grammatical differences. The data set for subsequent analyses and experiments consists of 303 pairs of user utterances, of which 88 pairs are associated with corrections of substitution misrecognition errors and 215 are tied to corrections of rejection errors.

Chapter 4

Acoustic Analysis

In the previous chapter we described in detail the environment in which the human-computer spoken correction data was collected. We explained the selection of 303 original input-repeat correction pairs, of which 88 were corrections of misrecognition errors (hereafter, CME's) and 215 were corrections of rejection errors (CRE's). In this chapter we will describe a group of acoustic analyses performed on these groups of utterance pairs. Specifically, we analyze these utterances under four broad classes of acoustic-prosodic features: duration, pause, fundamental frequency (f0), and amplitude. These measures draw from much of the literature discussed in chapter 2, but are based most heavily on those in [Oviatt *et al.*, 1996] and [Ostendorf *et al.*, 1996]. We will demonstrate significant differences between original input and repeat correction utterances in duration, pause, and fundamental frequency.

4.1 Duration

Duration has long been known to play an important role in a wide variety of speech phenomena. Ends of phrases and utterances are characterized by phrase-final lengthening [Allen *et al.*, 1987][1]. Final positions in lists are denoted by increased duration. ['t Hart *et al.*, 1990] Stressed and accented syllables are longer than those that are destressed or unstressed. [Nooteboom, 1997][2] Discourse segment- initial utterances also exhibit increases in duration relative to segment-internal utterances. [Swerts and Ostendorf, 1995] We will show that it also plays a significant role in spoken corrections.

For the majority of these analyses, the following technique was used to obtain utterance duration measures. A two-step semi-automatic process was required. First, the waveform and the corresponding utterance that had been segmented from the full conversational log were sent to a forced alignment procedure. The procedure used the Oregon Graduate Institute Center for Spoken Language Understanding (CSLU) CSLUsh tools [Colton, 1995] to produce a word-level forced alignment at a ten millisecond scale. A second pass over the automatic alignment was performed by a trained analyst. This pass was required to correct for any errors in the original alignment procedure; these errors arose from a variety of factors: background or

[1]Phrase-final lengthening is a phenomenon in which phoneme durations become elongated at the end of an utterance.

[2]The first syllable in 'teacher' is stressed; the second syllable is unstressed.

non-speech noise in the recording, pronunciation mismatched between the aligner dictionary and the spoken utterance, etc. The corrections focussed on three classes on position within the utterance: initial onset of speech, final speech position, and the boundaries of sentence-internal pauses. The goal was to delimit the total duration of the speech in a user turn, rather than to adjust all alignments. We took a conservative approach, only changing an alignment position if there was a better destination position available. From the alignments it was possible to automatically compute the following measures: total utterance duration, total speech duration, total pause duration, total number of pauses, and average length of pause.

4.1.1 Total Utterance Duration

The first measure we will consider is total utterance duration. Simply put, the total utterance duration is the length in milliseconds from the onset of the user speech in the utterance to the final speech position. Overall, utterances ranged in duration from 210 milliseconds to 5180 milliseconds. An example of an original-repeat pair with increase in total utterance duration appears in Figure 4-1. Analysis of Variance on duration and position (original vs. repeat) (Figure 4-2 yields $F(1,604) = 5.521$. (With log-transformed data, ANOVA yields $F(1,604) = 6.435$.) results yielded T-test two-tailed ($t = 1.97$, $df = 604$, $p < 0.05$) also indicates a significant increase in total utterance duration from original to correction utterances. Specifically, the mean length of an utterance is 864.1188 milliseconds for original input utterances and 969.0264 milliseconds for repeat correction utterances. This increases corresponds to a 12.15% increase in total utterance duration.

4.1.2 Total Speech Duration

Total speech duration calculates the difference between total utterance duration and total pause duration. This measures tries to capture the contribution of the speech segment, rather than an increases in number or length of pause, to the increase in total utterance duration, In other words, are users simply pausing more, lengthening phoneme, or increasing both pause and phoneme length. An example of an original-repeat correction pair in which speech duration increases with no corresponding increase in pause number or duration appears in Figure 4-4. Again analysis of variance for duration and position yields $F(1,604) = 4.52$ (Figure 4-3). (With log-transformed data the ANOVA yields $F(1,604) = 5.908$.) T-test two-tailed ($t = 2.17$, $df = 604$, $p < 0.05$) indicates an increase in speech duration from original to repeat inputs. This value corresponds to an average increase of 9.5%.

4.2 Pause

Pauses, the presence of unfilled silence regions within utterances, can play a significant role in discourse and utterance-level prosody. In discourse-neutral speech, pauses generally appear at intermediate and intonational phrase boundaries, which

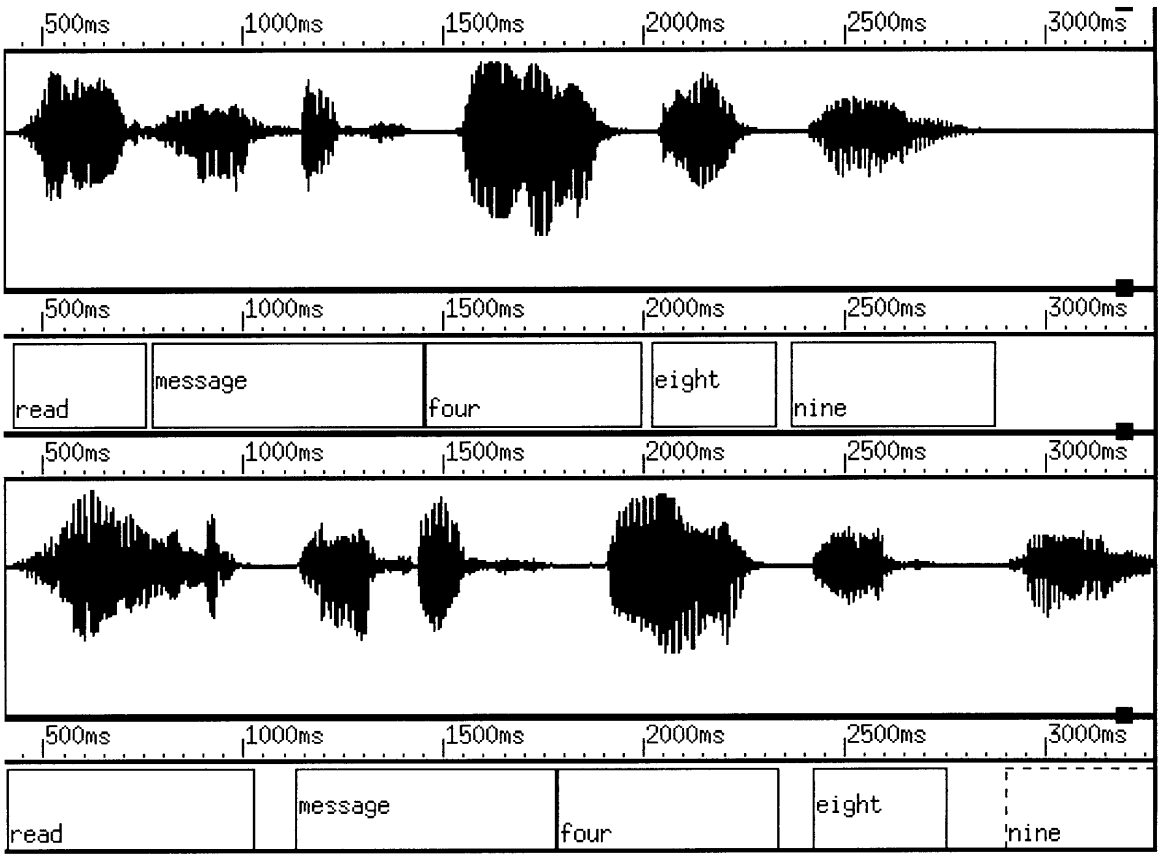


Figure 4-1: Original (top) - Repeat (bottom) pair with increase in total duration, pause duration, and speech duration

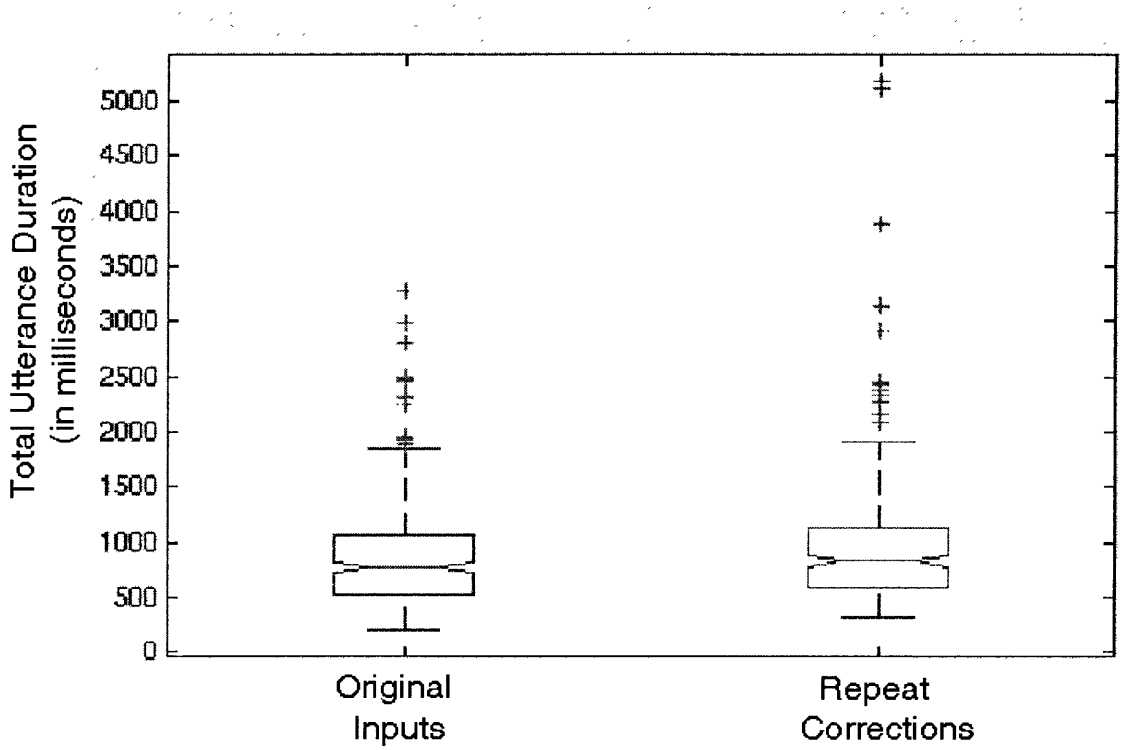


Figure 4-2: Original vs. Repeat Total utterance duration: There is an average 12.5% increase in total utterance duration between original inputs and repeat corrections.

often coincide with syntactic phrase or sentence boundaries. [Pierrehumbert, 1990], [Bachenko and Fitzpatrick, 1991] Speech systems commonly rely on extended periods of silence, one second or more in length, to identify the end of the user's turn. [Yankelovich *et al.*, 1995] While this method is arguably not a good way to detect turn transitions, it is, however, quite effective. The presence of lengthy pauses was found to be a strong cue to the start of a self-repair or other disfluency.[3] [Nakatani and Hirschberg, 1994], [Heeman and Allen, 1994], [Shriberg *et al.*, 1997] Pauses exceeding 50 milliseconds in length also proved useful in discriminating among speaking styles. [Ostendorf *et al.*, 1996]

Here, as noted in the discussion of duration measures, we coded the beginning and ending positions of all pauses in the original-repeat pair data. Silences were coded as pauses only if they exceeded 20 milliseconds in duration. In addition, we excluded all pauses prior to unvoiced plosives (k,t,p) and affricates (e.g. ch).[4] This choice was made due to the need to arbitrarily place the starting position of the unvoiced closure for phonemes of these classes, making it impossible to accurately determine the length or even existence of a preceding pause. For each utterance, we then computed the

[3]A disfluency is a disruption in normal speech. There are many types: pauses, 'filled pauses', where the speaker inserts 'um' or 'uh', or repetition, as in 'read the the message'.

[4]These phonemes are just a subset of the consonants where the vocal chords do not vibrate at the beginning of the sound. Since speech analysis tools depend heavily on this information, it is hard to identify the start of these sounds precisely.

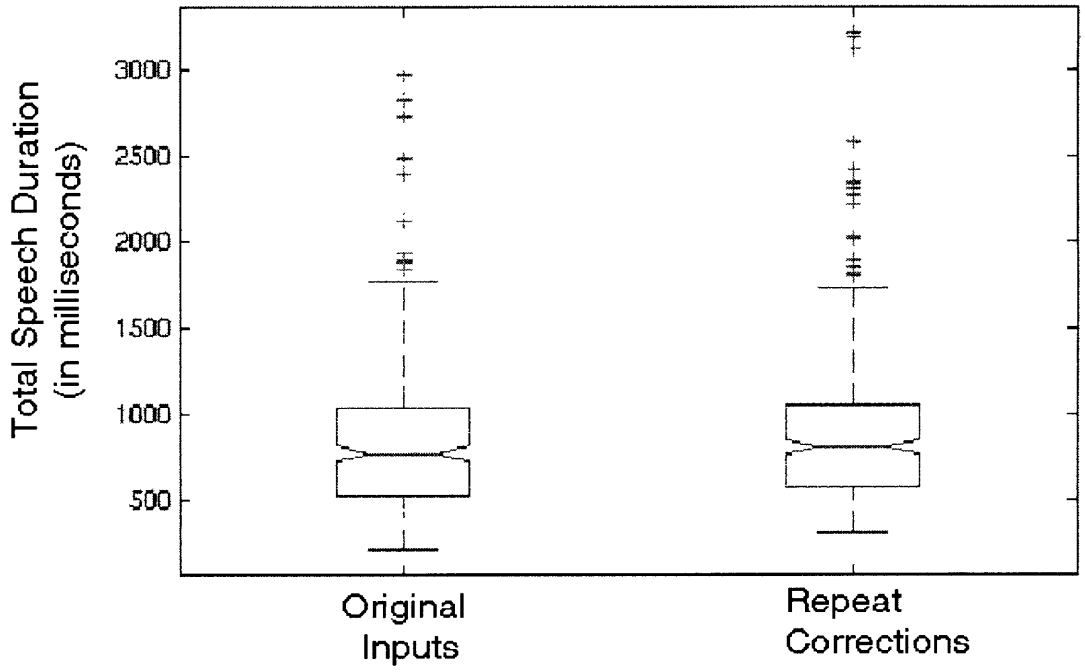


Figure 4-3: Original vs. Repeat Total speech duration: There is average 9.5% increase in speech duration, utterance duration excluding silence between original inputs and repeat corrections.

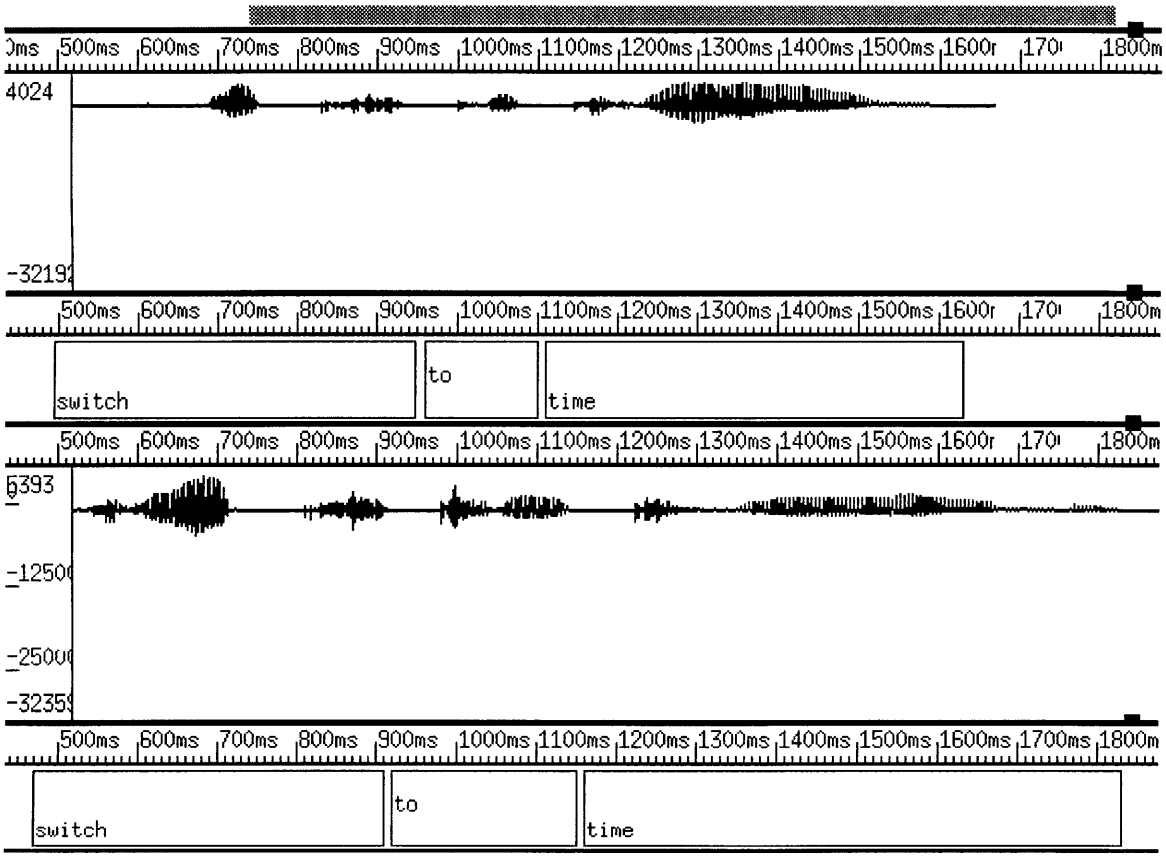


Figure 4-4: Original (top) vs. Repeat (bottom) pair with increase in speech duration only

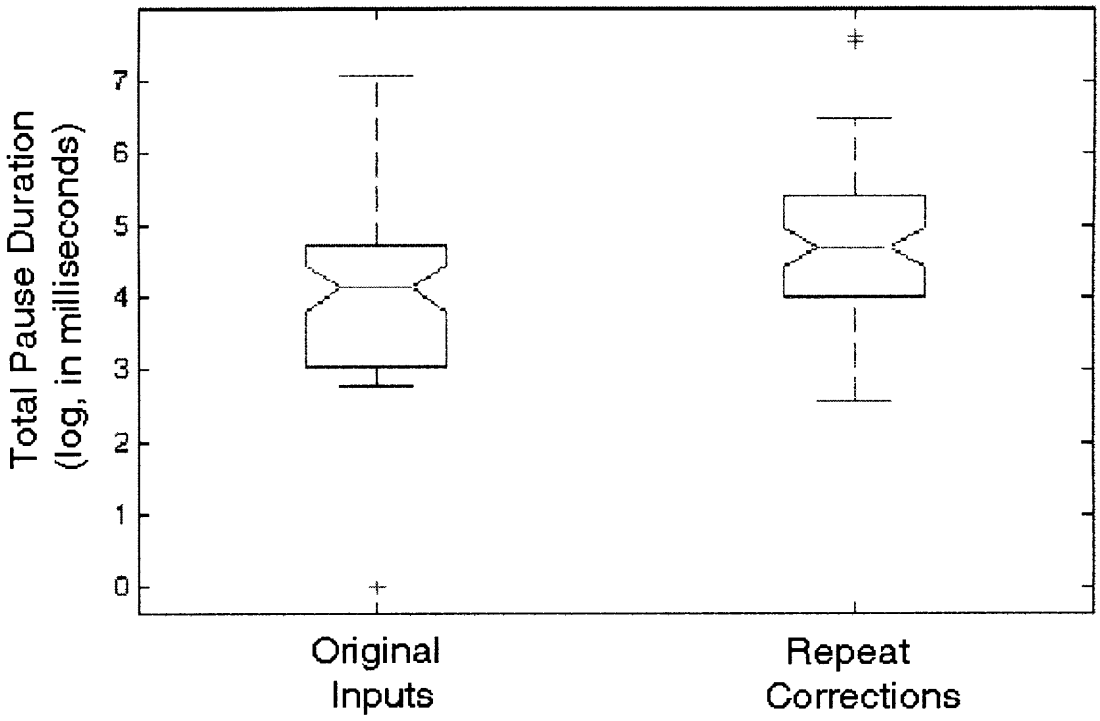


Figure 4-5: Original vs. Repeat Total Pause Duration: There is an average 59% increase in total pause duration between original inputs and repeat corrections.

length of each pause, the total number of pauses, and total pause duration. Figure 4-6 below illustrates an increase in pause number and duration with little increase in speech duration.

For all pause duration comparisons we considered only those utterances with at least one pause. Analysis of variance for pause duration of original versus repeat correction inputs gives $F(1,132) = 5.097$. (Figure 4-5 (With log-transformed data, ANOVA results in $F(1,132) = 15.94$). T-test, two-tailed, also yields significant results ($t = 2.2$, $df = 132$, $p < 0.05$) indicating an strong increase in pause duration. Specifically, within utterance silence regions increase from an average of 104.1791 milliseconds for original input utterances to an average of 165.0597 milliseconds, corresponding to an average increase of 59% in total pause duration.

Total utterance duration was tied to increase in pause duration. To measure these changes we computed the ratio of pause duration to total utterance duration for both original and repeat utterances where pauses occurred. We then performed analysis of variance on these ratio measures finding $F(1,132) = 5.2$. (Figure 4-7)(With log-transformed data, ANOVA produced $F(1,132) = 5.815$). T-test two-tailed also yielded significant results ($t = 2.28$, $df = 132$, $p < 0.025$) showing an increase in the proportion of silence to total utterance duration. From an average of 7.28% in original utterances, the proportion of silence increases to 10.56%, corresponding to an increase of 46% in the proportion of silence in an utterance.

We computed a final composite measure of speaking rate in number of syllables

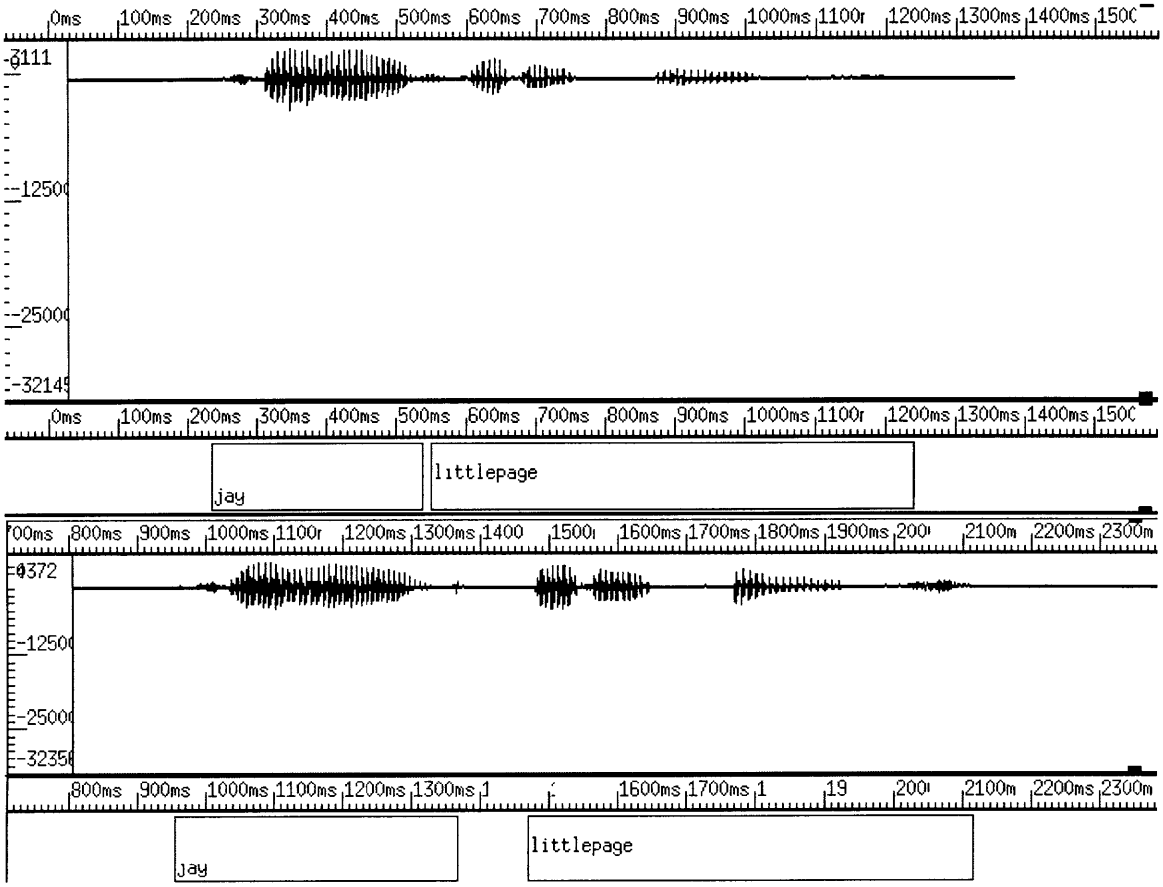


Figure 4-6: Original (top) - Repeat (bottom) pair with increase in pause duration: Note the insertion of silence between "Jay" and "Littlepage" with no additional increase in word durations.

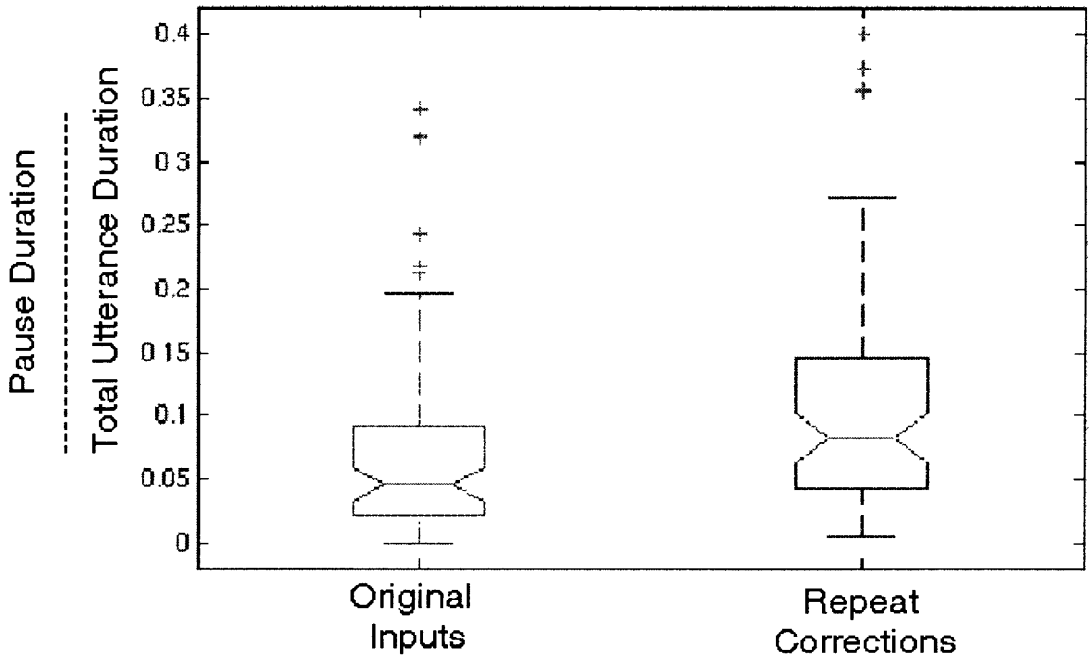


Figure 4-7: Original vs. Repeat: Ratio of Pause to Utterance Length: The proportion of silence in an utterance, relative to speech, increases an average of 46% from original inputs to repeat corrections. Both speech and silence duration increase in correction utterances, but silence increases more, proportionately.

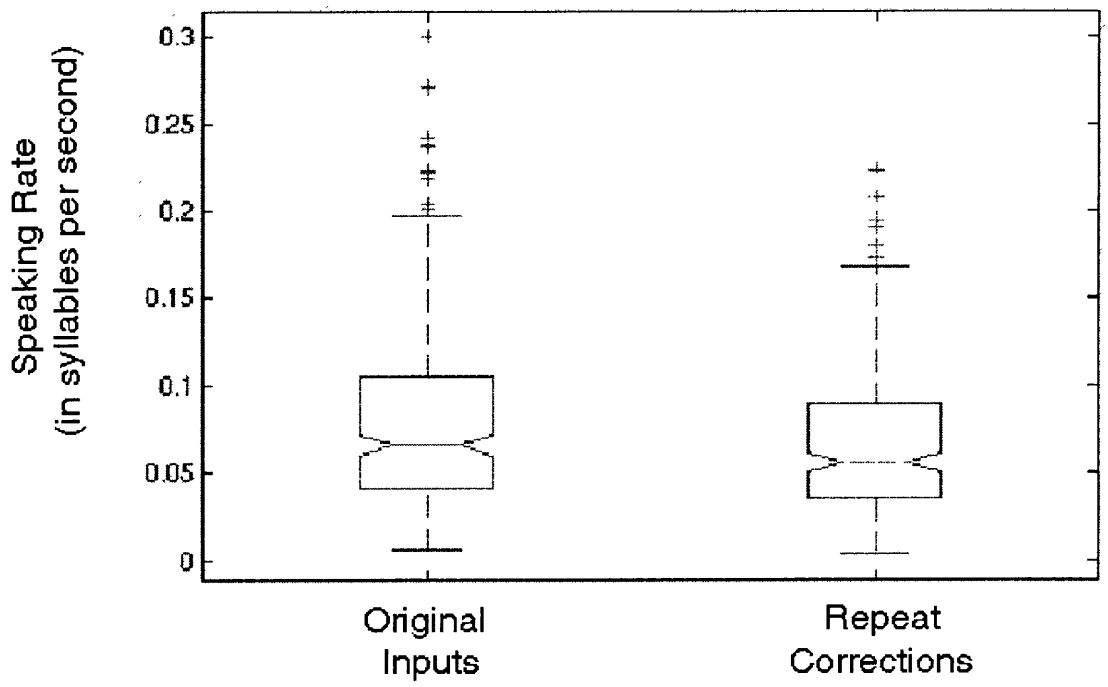


Figure 4-8: Original vs. Repeat Speaking Rate: Corresponding to the increases we have observed in speech and pause duration, we find that speaking rate, measured in syllables per second, decreases significantly (20%) from original inputs to repeat corrections.

per second and normalized by utterance duration. [5] Again looking at original and repeat correction utterances, we performed analysis of variance yielding $F(1,604) = 16.95$. T-test two-tailed ($t = 3.6$, $df = 604$, $p < 0.001$) demonstrates a significant *decrease* in speaking rate from original to repeat. The average speaking rate for original utterances was 0.0807 dropping to 0.0651 for repeat utterances, a decrease of 19.3%. (Figure 4-8)

4.3 Fundamental Frequency

Fundamental frequency (f0), pitch, presents a knottier problem than duration or pause, largely because, unlike the two preceding measures, pitch is not a simple scalar quantity.[6] Thus we must consider not only values but also rate and direction of change. Pitch plays an undeniable role in spoken language understanding. In addition to tone languages like Chinese, where pitch is part of the lexical identity of each word, pitch still plays a vital role in tune languages like English. Final rising or falling

[5]Measuring speaking rate in phonemes or syllables per second is a standard approach in speech systems, compensating for differences in content of the utterances that make measures like number of sentences per second unreliable.

[6]The most common instance of pitch contrast is between male and female speakers. Men generally have lower pitched voices, while women have higher voices.

contours can distinguish lexically and syntactically unmarked questions, declaratives, and commands. The degree of use of these contours also distinguishes conversational from read speech. [Daly and Zue, 1996] Pitch accents [7] and contours are used to distinguish given and new information [Terken, 1997] and indicate focussed information, As noted earlier, expanded pitch range has been associated with discourse segment-initial utterances. [Swerts and Ostendorf, 1995], [Nakatani *et al.*, 1995]

The basic coding of fundamental frequency is straight-forward. We used the ESPS/Waves+ signal processing package to compute the f0 for samples every 10 milliseconds throughout the utterance, taking only values where the Waves+ voicing detection reported positively. In addition, we excluded all points where RMS energy was less than 300, to avoid syllable onset and offset distortions.[8] Finally, we removed all erroneously doubled and halved pitch values that resulted from tracker error or from regions of extreme glottalization. [9] From these values, we computed maximum and minimum pitch values for each word and for the utterance as a whole. In addition, we noted the contour, rise, fall, or complex, of each word and the final position in the utterance in particular.

Now given that some of the subjects were female although a majority were male, it was necessary to normalize the absolute pitch values with respect to speaker. Thus for each subject we computed a pitch mean and standard deviation. From this base we compute a normalized set of pitch measures; these normalized values were computed as: $\frac{\texttt{pitchval - subjectpitchmean}}{\texttt{subjectpitchstddev}}$ for each of pitch maximum and minimum. Normalized pitch range was computed by $\frac{\texttt{pitchmax - pitchmin}}{\texttt{subjectpitchstddev}}$ This measure, like kurtosis, measures the length of the tails of the distribution of pitches. It allows us to identify compressed or expanded pitch ranges, the latter being identified with discouse function.

Finally, a last group of pitch measures was designed to capture a quantitative measure of the pitch contour. We derive a piecewise linear slope measure, computed by connecting the pitch maxima and minima of each word , dividing by the time between each peak. Results for these measures are shown in the following sections.

[7] Pitch accent is a particularly high or low pitch on the stressed syllable of a word; it usually appears on the main word of a phrase.

[8] Pitch tracking is only really accurate for vowels, where the waveform is approximately sinusoidal. The algorithms try to detect this voicing, but boundaries can pose problems.

[9] This phenomenon often occurs at the end of utterances. It results from a drop in lung pressure and relaxation of the vocal folds. These changes result in long individual glottal pulses perceived in the waveform. Importantly, these changes result in highly variable and abnormally low F0 measures, as low as 3-50 Hz, much lower than the base speaking pitch for any speaker.

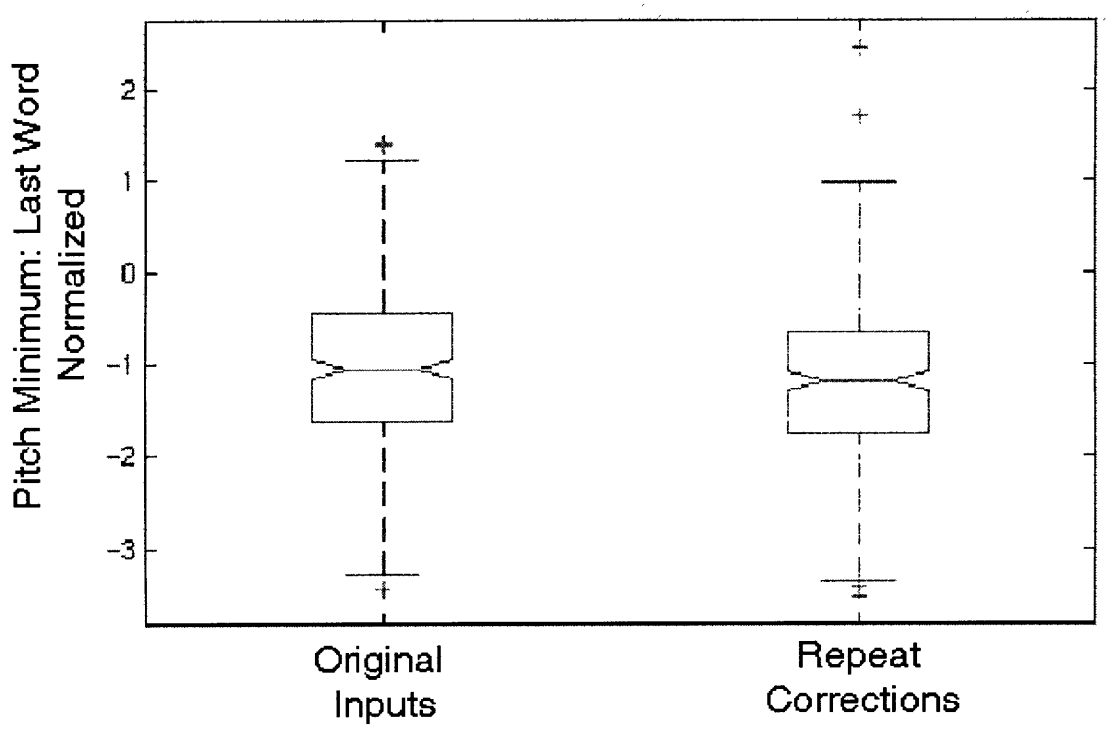


Figure 4-9: Original vs Repeat: Final Word Pitch Minimum: After normalizing for pitch differences between speakers, particularly male and female contrasts, we can identify a significant decrease in the lowest pitch of the final word in correction utterances in contrast to original inputs.

4.3.1 Scalar Pitch Measures

We found no significant differences between original inputs and repeat corrections for pitch maximum or pitch range (normalized or not normalized). However, we did find significant decreases in pitch minimum, both global to the sentence and for the last word in the sentence, when pitch was normalized and reported in terms of number of standard deviations from a per-speaker mean. Analysis of variance demonstrated a significant effect of position (original versus repeat) on pitch minimum for words in final position. ($F(1,604) = 3.963$) T-test showed a significant decrease in pitch minimum. (One-tailed, $t = 1.98$, $df = 604$, $p < 0.025$)

Unsurprisingly, similar, even clearer, results also hold for overall utterance minimum. Specifically, ANOVA yields a significant effect of original versus repeat correction position on global normalized pitch minimum. ($F(1,604) = 5.205$) T-test, two-tailed, again show significant decrease in global pitch minimum. ($t = 2.27$, $df = 604$, $p < 0.025$)

4.3.2 Pitch Contour Measures

Now we shift from static, scalar measures of pitch extrema to measures of pitch movement over time in terms of pitch contour. The first measure we analyze in this

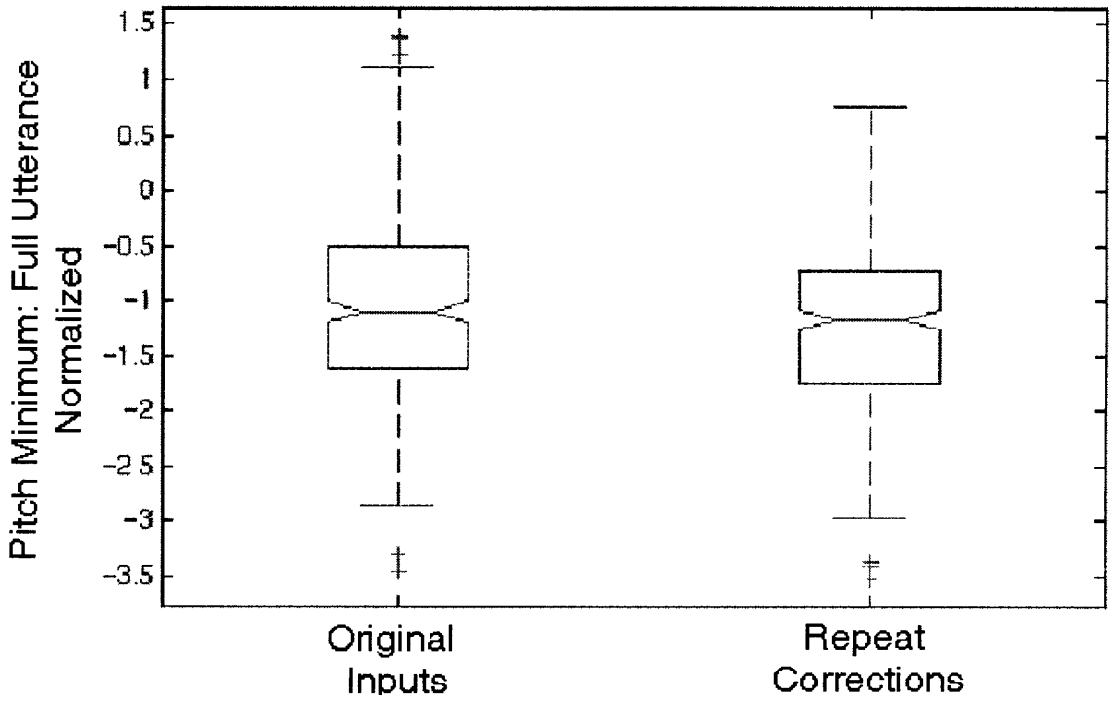


Figure 4-10: Original vs Repeat: Global Utterance Pitch Minimum: When we examine the lowest pitch for the full utterance, again after normalizing for inter-speaker variation, we find that the lowest pitch reached by speakers in correction utterances is significantly lower than that in original inputs.

section is a qualitative measure of the direction of the final contour of the utterance. This measure was coded as "Rise", "Fall", or "Other", where the contour was either complex or indeterminate. We found that for original inputs 70% exhibited a final falling contour, 18% a final rise, and 12.5% a complex or flat contour. We noted a marked contrast with repeat corrections where 77.6% of final contours were falling, 12.5% were rising, and 10.7% took on other contours. Almost one third of the rising contours change to falling contours, a very substantial shift. [10]

The final group of pitch measures are slope measures, rate of change of pitch over time, positive figures being rises and negative being falls. Now we have already dealt with the final contours in the immediately preceding analysis, so for these contour analyses, we exclude the final contour segments. This exclusion is also argued for on the grounds of pitch analysis. The final pitch segment corresponds to the boundary tone of the intonational phrase for the utterance. [11] Excluding the boundary tone allows us to examine the sentence-internal pitch and phrase accents without, possibly confounding, interference from boundary tone. While no pitch slope measures reached significance for original-repeat pairs in general, we find interesting effects when we treat the two classes of corrections, corrections of rejection errors and corrections of misrecognition errors, separately.

First we compare steepest slope rise for corrections of misrecognition errors for original inputs to that for repeat corrections. T-test, one-tailed, shows a significant increase in slope of the steepest rise ($t = 1.73$, $df = 124$, $p < 0.05$).

Likewise, there is a significant effect of original versus repeat position for sum of steepest rise and steepest fall slopes, though no effect for fall slope reached significance. ($F(1,174) = 3.98$) T-test, one-tailed, shows significant increases in the sum of slopes measure as well. ($t = 1.98$, $df = 174$, $p < 0.025$)

Next we consider corrections of rejection errors. Now none of the slope measures reached significance for this class of corrections alone. However, we also performed comparisons of corrections between the two classes. Here we find significant increases for slope rises (t-test, two-tailed, $t=2.7$, $df = 302$, $p < 0.01$) and slope sums (t-test, one-tailed, $t=1.69$, $df = 302$, $p < 0.05$) from corrections of rejection errors to corrections of misrecognition errors. An example of an increase in pitch accenting appears below. (Figure 4-12)

4.4 Amplitude

The amplitude or loudness of speech is associated with several important speech features. Increased amplitude, like pitch, is a characteristic of stressed syllables

[10] Some speakers take on idiosyncratic pitch contours in corrections. One subject consistently shifted to final rising contours, directly opposite to the overall group behavior. Another shifted to a rise on the first word of the utterance before continuing to a final fall. In general, we did not observe a shift to a 'list' style of speaking with a rise-fall on each lexical item associated with corrections. It did arise spontaneously in password entry, a digit sequence, and name entry for some speakers.

[11] Boundary tone is a specialized term that can be understood simply as the final pitch contour of the utterance.

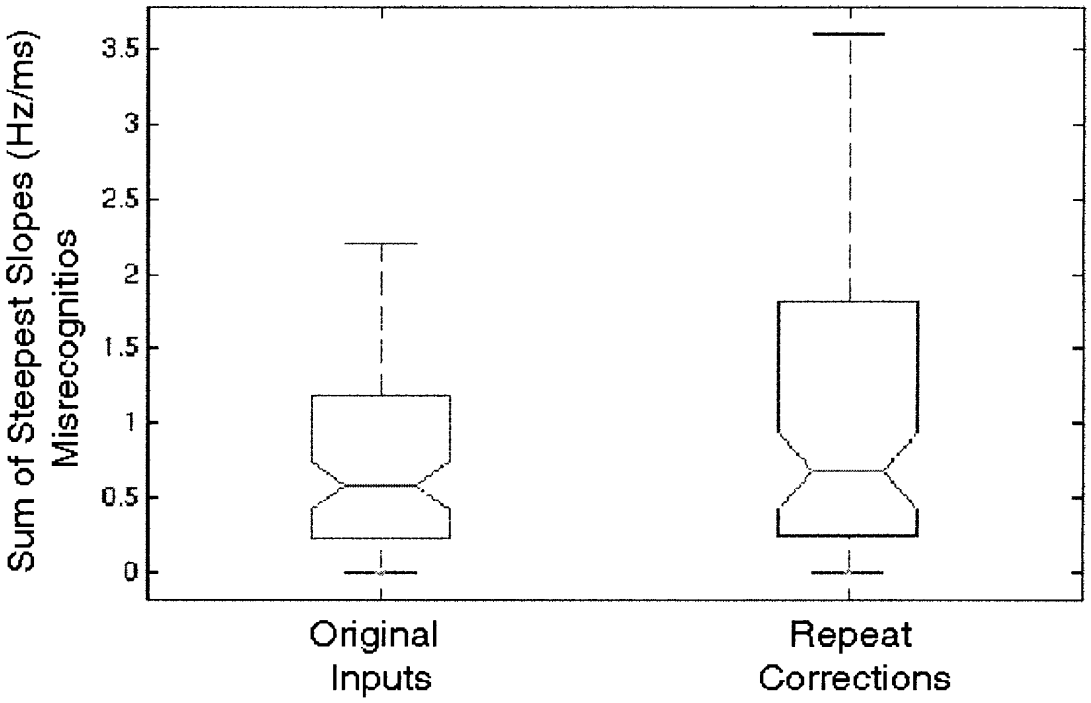


Figure 4-11: Original vs Repeat: Sum of Steepest Slopes: Misrecognitions: In order to capture a quantitative measure of the presence and magnitude of contrastive use of pitch accent, we compute a piecewise slope of the pitch track for the utterance. An utterance with strong pitch accent should have steep rises and falls, so we sum the slopes of the steepest rise and steepest fall in the utterance. Comparing this measure for original inputs and repeat corrections of misrecognition errors only, we find a significant increase in this measure of pitch variability for corrections.

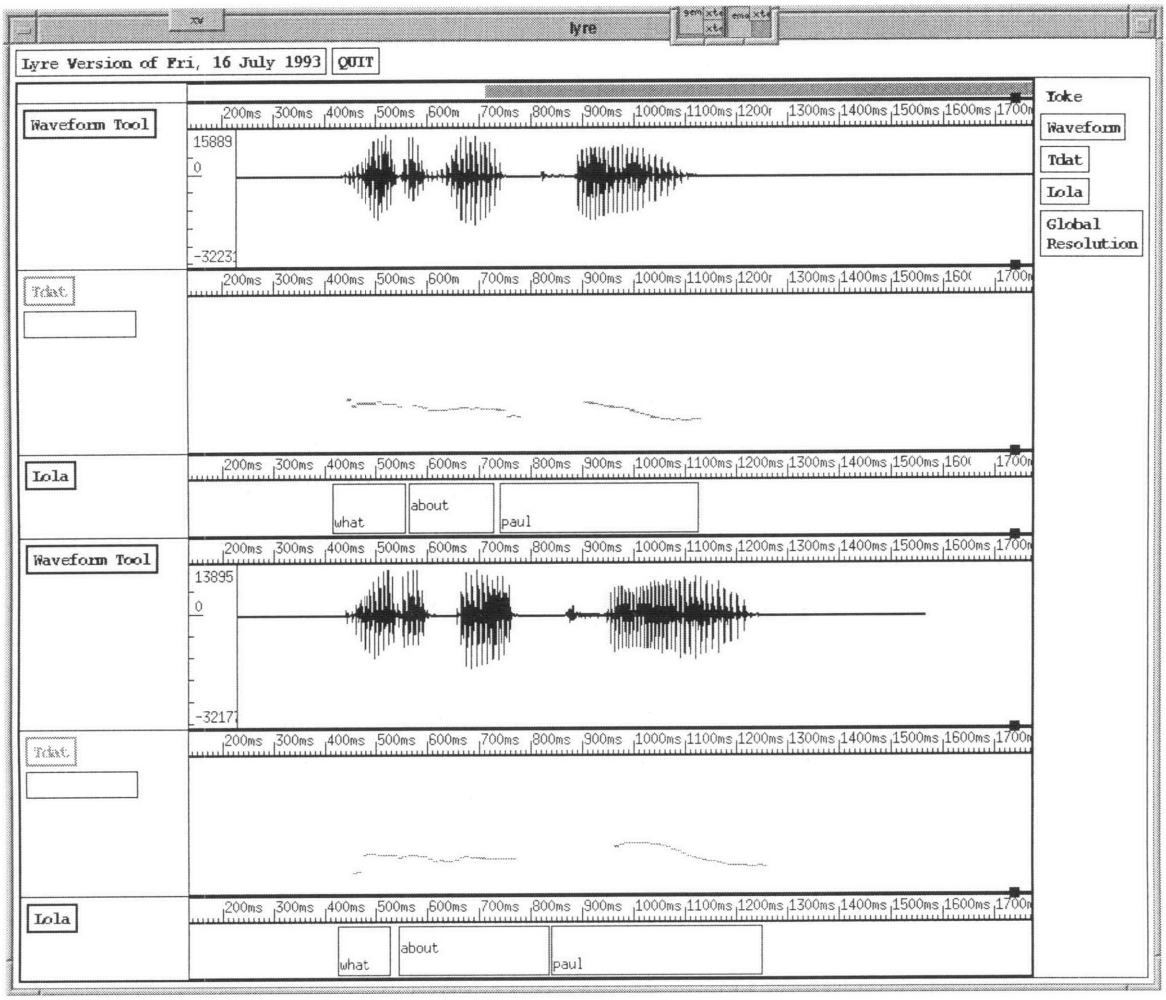


Figure 4-12: Original (top) - Repeat (bottom) pair: last word (Paul) changes from falling contour to high accent +fall: Unlike previous comparisons, this figure displays three lines for each utterance: the waveform (top), pitch track (middle), and test alignment (bottom). The important contrast is in the pitch track for the last word (Paul) in the two utterances. In the top (original) case, the pitch track is mostly level to a final falling contour, or low boundary tone. In the bottom utterance (repeat), there is a rise on the last word to a high pitch followed by a final fall to a low boundary tone, showing an increase in pitch accentuation from the original input.

and accented words, contrasting with the amplitude of surrounding words or syllables, Increases in amplitude have been linked to signaling self-repairs [Nakatani and Hirschberg, 1994]. A signal-to-noise ratio measure has also been used in experiments in disfluencies, self-repairs, and speaking style, and has been shown to be useful in detecting these phenomena. [Shriberg *et al.*, 1997]. [Ostendorf *et al.*, 1996]

In measuring amplitude in the following experiments, we used two measures computed by the ESPS/Waves+ signal processing system: log raw power and RMS energy.[12] A quick verification pass was made by a trained analyst to exclude any values contributed by background or non-speech noise in the recording. Next, automatic procedures computed several derived amplitude measures, For log raw power measures, maximum and average amplitude values were computed for all utterance regions above 30dB, to avoid lowering the amplitude measure because of silence regions. For RMS energy measures, values from all voiced regions (again relying on the ESPS/Waves+ voicing detector) were used to compute average values, while the maximum value was smoothed across the three highest RMS energy values. The same methods were used to compute maximum and average amplitude values for each word region.

As with pitch, amplitude is highly variable across subjects and interactions. We again computed a per-subject normalization term based on the mean and standard deviation of each amplitude measure. With these terms we also compute derived amplitude measures:

$$\frac{\texttt{amplitude - subjectampmean}}{\texttt{subjectampstddev}}$$

Although amplitude is, anecdotally, one of the features commonly associated with corrections, we found that none of the amplitude measures, normalized or not, reached significance.

4.5 Discussion

We have examined a variety of acoustic-prosodic measures contrasting original input and repeat correction utterances. These measures fall into four broad categories: duration, pause, pitch, and amplitude. We found significant differences between original and repeat utterances for all classes of measure, except amplitude. The discussion below will present a unified account of these changes.

4.5.1 Duration and Pause: Conversational-to-(Hyper)Clear Speech

Duration and pause measures will be considered together. We found significant increases in total utterance duration, total speech duration, total pause duration, speaking rate in syllables per second, length per pause, and proportion of silence in utterances between original inputs and repeat corrections. In correction utterances, users

[12] In speech, one often uses log measures to approximate the *perceived* increase in loudness of sound.

speak more slowly both by increasing the duration of phonemes within the utterance and by inserting or lengthening silence regions within the utterance. These changes fit smoothly into an analysis of corrections as shifting from more conversational, casual speech to more clear or careful speech along the continuum.

These contrasts are very similar to those in [Oviatt *et al.*, 1996], even to the extent that they reflect the same percentage increase in total utterance duration and total speech duration. The changes also echo those reported in [Ostendorf *et al.*, 1996] for contrasts between conversational and read speech; although precise figures for durational and pause change are not reported, they find increases in phoneme duration, presence of pause, and decrease in speaking rate to be correlates of more formal or read speech in contrast to conversational speech.

This increase in duration seems to be the most robust clear speech attribute. Other types of speech associated with clear style, such as speech to the hearing-impaired or speech to children (motherese)[13], exhibit increases in duration. On one hand, speech to children is often associated with higher pitch and expanded pitch range, while speech to the hearing-impaired lacks those pitch features but is associated with significant increases in loudness, [Fernald *et al.*, 1989], [Picheny *et al.*, 1986] We have noted distinctive pitch phenomena associated with corrections as well which are not shared by other clear speech styles. Speaking rate thus stands out as the most consistent clear speech feature.

One contrast with the [Oviatt *et al.*, 1996] analysis is that, while there is a significant increase in pause duration in both sets of data, that observed for the SpeechActs data is not as that observed in other work. Presence or absence of a 50 millisecond pause is not a deciding contrast between original and repeat correction as it is for the classes studied by [Ostendorf *et al.*, 1996].[14] A likely reason for this observed contrast is the length of the utterances in the current study. In the SpeechActs data overall, the average length of an utterance is between two and three words, and the average analyzed utterance duration is under two seconds. For the SpeechActs data, none of the analyzed utterances exceeded 10 words, while the data in [Oviatt *et al.*, 1996] includes sixteen-digit strings representing credit card numbers. Systems which predict pause location and prosodic phrasing typically use a combination of syntactic phrase structure and number of words or syllables in determining pause placement. Thus pauses are not distributed uniformly over utterances, but are unlikely to appear at all in very brief utterances. The sentences in the SpeechActs data are short enough to discourage pausing, creating this contrast in pause lengths.

4.5.2 Pitch: Accent, Contour, and Contrast

While we also found significant differences in several pitch measures between original inputs and repeat corrections, these contrasts, unlike those for duration and pause,

[13]Motherese refers to characteristic speech of caretakers to children. It is found in many languages, though more for females than males, and involves expanded pitch range, higher pitch, and longer duration.

[14]Presence of a larger pause duration (70ms) or larger proportion of silence does play a secondary role in classifying rejection errors with only acoustic information.

do not fit smoothly into an analysis of corrections as a uniform transition from a more conversational to a more clear speaking style. We will divide this section into discussion of pitch changes that are consistent across correction classes and pitch changes that differ significantly depending on the type of correction being performed.

One trend we find that holds for original-repeat pairs throughout the SpeechActs data is that of a decrease in subject-normalized pitch minimum. In other words, the lowest fundamental frequency reached in an utterance is lower for repeat corrections that for original inputs. Two possible explanations lend themselves to the cause of this drop in pitch minimum; this drop holds for the sentence overall and for the final word in the sentence. One explanation is the phenomenon of systematic downstepping of pitch and amplitude throughout the sentence. This downstepping when combined with the durational increases reported earlier could lead to larger downstep effects and thus lower pitch minima.

The other possible explanation is linked to the other systematically observed trend, for final pitch contours to shift from rising or complex to falling contours. This flip in final pitch contour to a low final boundary tone would increase th number of falling pitch regions and thus lower the minimum pitch achieved in many utterances.

Interestingly, compared to other related work examined earlier in this thesis, only one study found similar phenomena, the most closely related study by [Oviatt *et al.*, 1996]. Studies of self-repairs generally found point-to-point increases in pitch. Studies of discourse structure and intonation generally found increases in pitch or pitch range rather than decreases. Nor did the read/conversational contrasted speech in [Ostendorf *et al.*, 1996] show any effects of pitch. Only one study by [Daly and Zue, 1996] makes note of differences in rising pitch contours in yes/no questions in read versus spontaneous speech. Daly observes that conversational speech exhibits more rising contours in yes/no questions, where theory generally predicts there should be a final rising contour, than does read speech. However, the rising contours in original inputs in the SpeechActs interactions do not appear to be related to yes/no questions. In fact, the SpeechActs system functionality does not enable any literal yes/no questions; most legal utterances are information-seeking questions or commands. Instead, since these utterances should canonically have falling intonation, we can view these changes as shifts from a casual or tentative style to something closer to citation form.

It has been suggested that corrections are, de facto, segment-initial utterances since their use initiates a clarification subdialog. [Swerts and Ostendorf, 1995], citeG-H However, we find little prosodic evidence to support this analysis. The most salient feature of discourse segment-initial utterances, as identified by many researchers including [Swerts and Ostendorf, 1995], [Nakatani and Hirschberg, 1994], citeG-H, is a significant increase in pitch range, often linked with an increase in pitch height. In our analysis of corrections, on the other hand, we find no significant expansion in overall pitch range or increase in pitch height; conversely, we find overall decreases in pitch minimum for these utterances. This contrast argues that, at least for this simple type of correction interaction, corrections do not have the same status as segment-initial utterances.

The final set of pitch contrasts we found involved a contrast between original inputs and repeat corrections of misrecognition errors, also in contrast to corrections of

rejection errors. Here we observe that there is greater pitch movement corresponding to steeper rises and steeper falls for corrections of misrecognition errors than for original inputs or repeat corrections of other types. These contrasts indicate the presence of stronger or new contrastive pitch accents. This specific use of contrastive accent would not be expected to appear in general contrasts between original inputs and repeat corrections or read and conversational speech. It would not necessarily be a part of intonational marking of discourse structure, except in cases of parallel contrastive structures. [Prevost, 1996] This type of accentuation would be most likely to appear in self-repairs where we observe pitch increases between reparandum and repair. This increase in pitch activity appears to be tied to accenting, rather than to any overall contour and would thus be difficult for [Taylor, 1995] to detect.

Chapter 5

Decision Tree Classification

The analyses in the previous chapter provided evidence of significant differences between original input and repeat correction utterances, surfacing in a number of different acoustic-prosodic features. Specifically, correction utterances were shown to be significantly longer in total utterance, total speech, and total pause duration than original inputs. In addition, correction utterances exhibited significant decreases in pitch minimum. In this chapter we design decision tree classifiers to distinguish between original inputs and repeat corrections for corrections of rejection errors and corrections of misrecognition errors. Such a classifier would be incorporated in a spoken language system help defuse error spirals, by identifying corrections and initiating repair interactions.

5.1 Decision Trees: Motivation

In order to correctly interpret the utterance and also, if necessary, to invoke a specialized recognizer to handle the adaptations involved in a correction utterance, our initial goal is to be able to distinguish between original inputs and repeat corrections. For this classification task we have a variety of duration, pause, pitch, and amplitude features at our disposal. From our prior statistical analysis of these acoustic-prosodic features, we can presume that some, probably duration and pitch, will be more useful than others, such as amplitude. Still, it is not possible to be certain which features will be most effective, and it would be most informative to be able to determine by inspection of the most successful classifier, exactly which features or sets of features contributed to this success.

Decision trees provide a machine learning technique that can fulfill these requirements. First, decision trees can ignore irrelevant attributes. In nearest neighbor classification techniques, for example, each training and, later, testing instance can be viewed as a labeled point in a high dimensional space, where each feature value corresponds to a dimension in the space. Thus, all features carry equal weight in the classification process, and the inclusion of irrelevant features can cause otherwise similar instances to become widely separated in the classification space. In contrast, decision trees make selective use of the most relevant attributes and can therefore ignore irrelevant attributes. They achieve this behavior in the following way. The decision tree consists of several layers of branching nodes. Each of the branches corresponds to a split on a feature value, e.g. greater than or less than for a continuous

real-valued feature or one branch per possible value for an enumerated feature. At each stage the best split is chosen, where the split creates the lowest total entropy of the branches after the split. The aim is to create the purest possible clusters from the split. All possible assignments for all possible features are evaluated. Thus, a feature is only used for a branch decision if it yields subtrees with the lowest possible entropy at that point. While such a greedy heuristic may be misled, it will not select features which lead to highly heterogeneous branches. It must always select the test at any point that gives the greatest improvement in homogeneity. Finally, the leaf nodes receive a classifier label. To label a new unseen instance, one simply traverses the tree taking the branch dictated by the associated test at each point, giving it the same label as the other instances sharing its leaf node. For instance, a trivial possible classifier for corrections could label leaves as original or correction. Suppose that the root node tested whether the duration was greater than that of the previous utterance. A second test on both branches might be whether the utterance was louder than the previous utterance. Thus one might expect the following path through the tree: if the utterance is longer than the previous input and if the utterance is louder than the previous input, the utterance is a correction.

This methodology also gives rise to the other desirable feature of decision trees: perspicuous classification. By writing down each branch test for each path from root to leaf, one creates an easily intelligible set of if-then rules which describe the classification process. This intelligibility allows the designer to determine which features play the most important roles in the classification process. In contrast, techniques like nearest neighbor classification or neural networks are often very difficult to interpret. Nearest neighbor simply defines a collection of regions in the feature space within which a classification applies, giving an indivisible set of feature values or value ranges associated with that classification. Neural networks when trained produce a set of weights on different inter-node connections. With the exception of very low weights, near zero, and very high weights, little can be determined about which features and values play a role the classification output by the network.

So, in order to obtain classifiers that are relatively robust to irrelevant attributes and that could easily be interpreted, we chose to build decision trees to distinguish between original inputs and repeat corrections. However, there are several other machine learning methods that could be applied to this task in future work, and that have desirable characteristics. For instance, decision trees define rectangular decision boundaries that may not be the best fit to the data attributes; they also treat all features independently. Bayesian techniques or mixture-of-experts approaches would allow different decision region shapes and could more effectively model independence and interdependence among features.

5.2 Classifier Features

Even though decision tree classifiers are fairly robust to irrelevant attributes, we still prefer to use features that are more likely to allow us to distinguish between original inputs and repeat corrections. Thus, we select features based on those that proved,

under statistical analysis, to exhibit significant differences between originals and repeat corrections. Therefore, we use *duration*, *pause*, *pitch*, and *amplitude* features. We will now describe, in detail, the features used in the decision tree classifiers and explain the basic analysis measures required to effectively use those decision trees.

5.2.1 Duration-related Features

In the acoustic analysis chapter, we noted the importance for pitch and amplitude measures of per-subject normalization. For duration, such normalization was not particularly important since there were systematic increases in duration. However, utterances range in duration from anywhere between 210 and 5180 milliseconds; this contrast depends primarily on the lexical content of the utterance. This variability makes original-repeat distinction based on absolute utterance length unlikely. As a result, in addition to using the absolute utterance duration, we experimented with a variety of normalization measures. One normalization measure referred to as *expected utterance length* is an average original duration calculated for each utterance text. This normalizing term was used in three different measures:

`lenvexp`	$\frac{\texttt{total utterance length}}{\texttt{average utterance length}}$
`lenvexpvowel`	$\frac{\texttt{total utterance length - average utterance length}}{\texttt{average utterance length}}$
`lenvexpsyll`	$\frac{\texttt{total utterance length - average utterance length}}{\texttt{number of syllables in utterance}}$

A rate of speech measure was also calculated from the number of syllables in the utterance divided by the square of the total utterance duration in seconds. This measure is a variant of the standard "syllable per second" speaking rate measure used in speech research; the second division by duration is motivated by the observation by [Ostendorf *et al.*, 1996] that such a repeated division improved performance in a similar classification task. Another set of normalized duration measures are based on a duration measure DDUR described in [Ostendorf *et al.*, 1996]. These measures are based on the following equation:

$$\sum_{w \in words} \frac{\texttt{actuallength(w)}}{\texttt{total utterance duration}} - \frac{\texttt{averagelength(w)}}{\texttt{total utterance duration}}$$ where `actuallength(w)` is the observed duration of word w and `averagelength(w)` is the mean length of the word w, calculated as the sum of the mean lengths of its constituent phonemes. We produced three features based on this measure: `ddur2pos` , the sum of all instances where `actuallength(w)` exceeds `averagelength(w)`, `ddur2neg`, the sum of all instances where `averagelength(w)` exceeds `actuallength(w)`, and `ddur2diff`, sum of the absolute values of all differences. These measures try to capture a fine-grained assessment of the divergence between the observed utterance durations and

the durations predicted by the model. One expects there to be significant increases in duration, but they might not be uniformly distributed among words. For instance, a stressed word might undergo a large increase in duration while function words such as 'the' change very little.

A third set of duration measures tries to capture the idea that a correction could alter the length of words throughout the utterance, decreasing the proportion of the utterance duration accounted for by the longest word and increasing the proportion of the utterance duration accounted for by the shortest word. Specifically, we have

$$\texttt{maxprop} \quad \frac{\texttt{duration of longest word}}{\texttt{total utterance duration}}$$

$$\texttt{minprop} \quad \frac{\texttt{duration of shortest word}}{\texttt{total utterance duration}}$$

In general, the best normalizations make use of the duration for a specific utterance text. Speaking rate measures whether based on observed syllables per second or acoustically determined phoneme or syllable rate, while not reaching the accuracy of the best text-based measures, perform fairly consistently. All of these measures aim at capturing the increased duration observed in repeat correction utterances in contrast to original inputs; they they simply perform this task and account for differences related to utterance content in different ways.

5.2.2 Pause Features

Pause duration also exhibited a significant difference between original and repeat correction in earlier acoustic analyses. This contrast also suggests that pause features could prove useful in building automatic classifiers for these speech acts, such as seen in speaking style discrimination [Ostendorf *et al.*, 1996] and self-repair identification. [Nakatani and Hirschberg, 1994], [Shriberg *et al.*, 1997]. A set of four measures were used to capture pause contrasts between original inputs and repeat corrections. One simple measure, `pausedur`, corresponds to the total pause duration. Another measure, `lenperpause`, is the average pause duration, $\frac{\texttt{total pause duration}}{\texttt{number of pauses}}$. `Pausenumperwd` computes the total pause duration divided by the total number of words in the utterance. A final measure, `pausevtotal`, captures the proportion of the utterance which is silence, $\frac{\texttt{total pause duration}}{\texttt{total utterance duration}}$.The first two measures provide absolute pause measures, while the last two normalize the pause length relative to different utterance length measures.

5.2.3 Pitch Measures

We used a battery of pitch measures for classifier design, in order to capture the different types of features, absolute extreme pitch value and measures of pitch slope and contour. As noted in the acoustic analysis, absolute measures of pitch are highly variable, particularly based on gender and subject. Therefore, numeric pitch values are presented in terms of subject-based standard deviation from a subject-based mean. Approximately half the measures relate to numeric pitch values, while the other half are measures of pitch slope or contour.

The numeric pitch values are measures of pitch maxima, pitch minima, and pitch range. Three measures are global pitch maximum, global pitch minimum, and global pitch range, calculated across voiced regions over the full length of the utterance. Two other measures are pitch maximum and pitch minimum for the last word in the utterance. These measures are based on those used in [Nakatani and Hirschberg, 1994], [Oviatt *et al.*, 1996], and [Swerts and Ostendorf, 1995]. These pitch extreme values capture pitch range expansion and height. They thus are linked to discourse structure, such segment beginnings, and it has been suggested that corrections fulfill such a role. They are also linked with certain clear speech styles such as motherese.[1] All of these measures are expressed absolute terms, in Hz.

There are six contour and slope-based pitch measures. One measure is `pitchdir`, the shape of the final pitch contour of the utterance; these values are either "Rise" or "Fall". `Pitchslope` is computed as $\frac{\texttt{(globalpitchmax - globalpitchmin)}}{\texttt{total utterance duration}}$. `Firstslope` measures the contour of the first word in the utterance; this term is $\frac{\texttt{(firstpitchmax-firstpitchmin)}}{\texttt{first word duration}}$. The other three pitch measures are measures of the peak-to-peak slope of the pitch contour, computed in piecewise linear fashion from consecutive pitch extrema. These measures are `maxslope`, the value of the steepest slope rise, `minslope`, the value of the steepest falling slope segment, and `slopesum`, the sum of the magnitudes of the steepest rise and steepest fall in the utterance. These measures of pitch slope were found to reflect significant differences between original inputs and repeat corrections for corrections of misrecognition errors. These slope measures aim to capture pitch accent behavior, such as contrastive accent, an expected component of corrections of misrecognitions.

5.2.4 Amplitude Measures

The amplitude measures used in the classifier essentially parallel the pitch measures. Again, to compensate for very high inter-speaker variability, we normalized amplitude measures based on a per-speaker mean amplitude, in addition to using absolute

[1] Motherese refers to characteristic speech of caretakers to children. It is found in many languages, though more for females than males, and involves expanded pitch range, higher pitch, and longer duration.

measures for these values. The base measures can be computed from either of the two amplitude measures described in the acoustic analysis chapter: log raw power and RMS energy. A group of six amplitude measures are computed.

Three measures are used to determine overall utterance amplitude. First we compute the utterance mean and maximum amplitude. We also compute the maximum amplitude of the last word in the utterance; this measure is a proxy for the sustained amplitude of the utterance. Next we compute three additional measures to try to capture the amplitude variability of the given utterance. `Ampdev` is the standard deviation of the amplitude for the given utterance. `Ampdiff` calculates the difference in amplitude from the beginning to the end of the utterance, specifically the difference in amplitude of the first and last words of the utterance. Finally, `ampdelta` represents the difference of the amplitude of the last word in the utterance to the mean amplitude of the utterance as a whole. These contrast measures consider whether the utterances follow common trends of *catathesis*[2], or systematic downstepping in pitch and amplitude from left to right through the utterance, or whether the greater articulatory effort found in clear speech or corrective utterances could override such a trend. Simply put, we hypothesized that in the more careful speech of corrective utterances users would make an effort to speak more consistently loudly, particularly in cases of corrections of rejection errors.

5.2.5 Feature Summary

Feature Class	Feature Names
Duration	Uttdur, Lenvexp, Lenvexpvowel, Lenvexpsyll Syllrate, Ddur2pos, Ddur2neg, Ddur2diff Minpos, Maxpos, Mrate
Pause	Pausedur, Lenperpause, Pausevlen
Pitch	Pitchmax, Pitchmin, Pitchrange Pitchlastmin, Pitchlastmax, Pitchslope, Firstslope Pitchdir, Maxslope, Minslope, Slopesum
Amplitude	Ampmax, Ampmean, Amplast Ampdev, Ampdiff, Ampdelta
General	SubjectID

5.3 Classifier Experiments & Results

Given the basic feature set described in the preceding section, we examined the use of decision tree classifiers for identifying original inputs in contrast to repeat corrections. We will contrast results based on the availability of different types of information to the classifier and results for different error correction types, corrections of rejection errors and corrections of misrecognition errors, since we observed in the acoustic

[2]This trend of decreasing pitch and loudness through the utterance is a result of decreasing lung pressure as one lets out one's breath in speech.

analysis presented in Chapter 3 that different correction types pattern differently for pitch features.

5.3.1 General Methodology

Each of the decision tree experiments presented below followed the same basic approach. The training and test sets were divided evenly between original inputs and repeat corrections; this division established a 50% baseline for recognition since simply guessing either classification for all cases would correctly classifying 50% of the instances. The leaves of the decision tree were labeled as either O(riginal) or R(epeat correction).

To avoid overfitting the training data, we evaluated the classifiers through 7-way cross-validation, operating in the following manner. We divided the data into seven segments, training on 6/7's of the data and testing on the remaining 1/7. The test set is drawn randomly according to the same distribution as the full data set. We then cycle through the segments so that every instance appears in the test set once. We report the results as the average of the accuracy rates over each of the test sets.

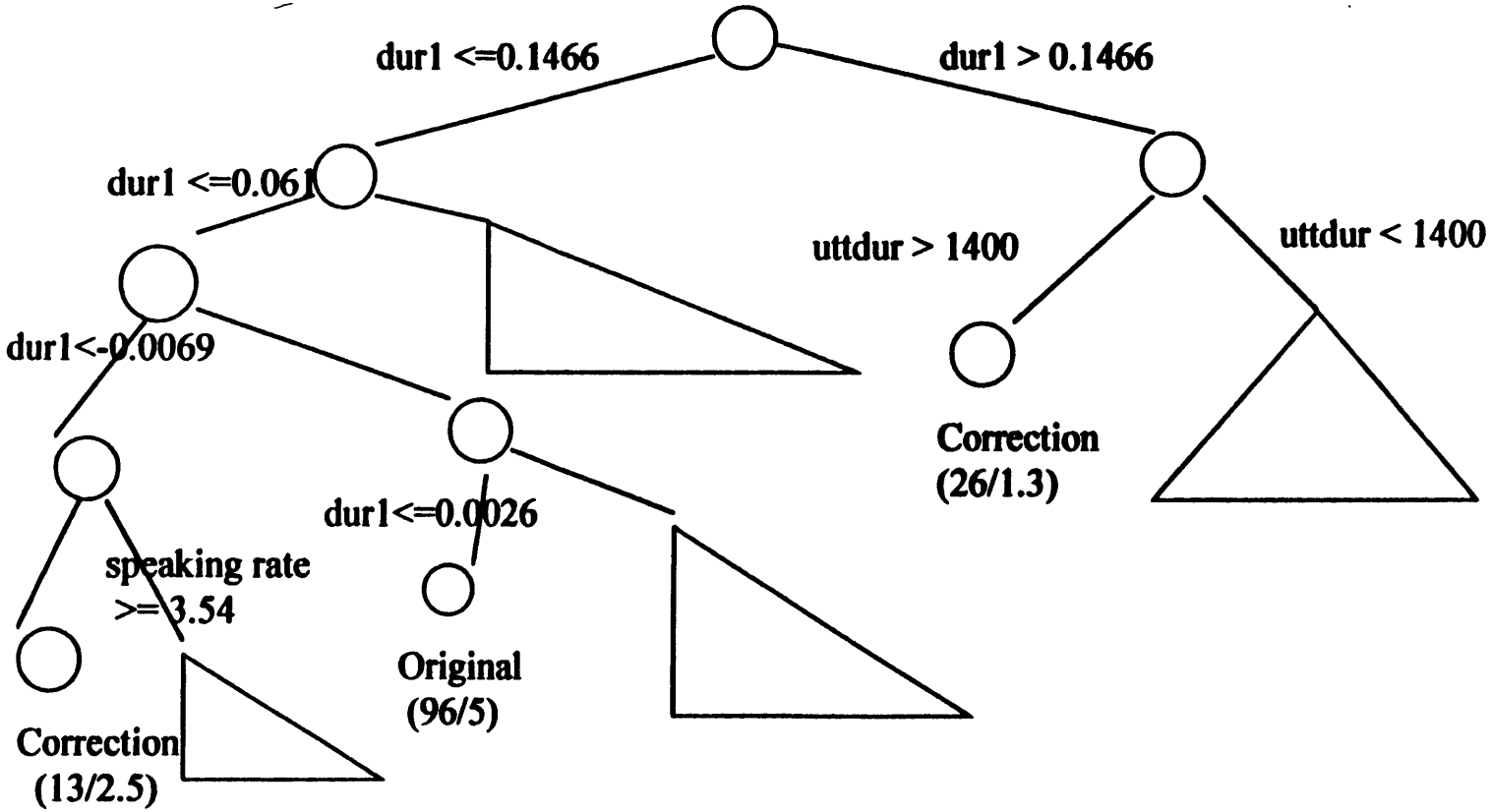
5.3.2 Recognizing Corrections: All Types

We begin by using decision tree classifiers to recognize corrections of all types in contrasted against original inputs. This classifier achieved 65% accuracy: in other words, a 35% error rate, in cross-validation tests for unseen data. The training set error for the same trees averaged 17%, or 83% accuracy. This rate falls between the classifier chance baseline of 50%, and a 79.4% baseline for human performance on a task where listeners were given utterances in isolation and asked to identify whether or not they thought a correction was taking place. [Rudnicky and Hauptmann, 1990] The best classifiers are all based on durational measures, specifically a combination of some form of normalized duration, absolute duration, and pause measures. A typical successful tree with 20-30 nodes required for a sensible split appears below. In all cases the first split in the classifier tree was based on a measure of normalized duration. This result is robust for several normalized duration measures. Specifically, we get similar results for `lenvexpvowel`, duration normalized based on average expected original input length for a given input text, as described the feature section above. We also get similar results for speaking rate computed as either syllables per second, $\frac{\texttt{number of syllables} * 1000}{\texttt{total utterance duration}}$ normalized per-subject, and from a pure acoustic measure called *mrate*, that automatically estimates the number of phonemes per second based on the number of spectral peaks per time period, developed by [Mirgafiori *et al.*, 1995] at Berkeley.[3]

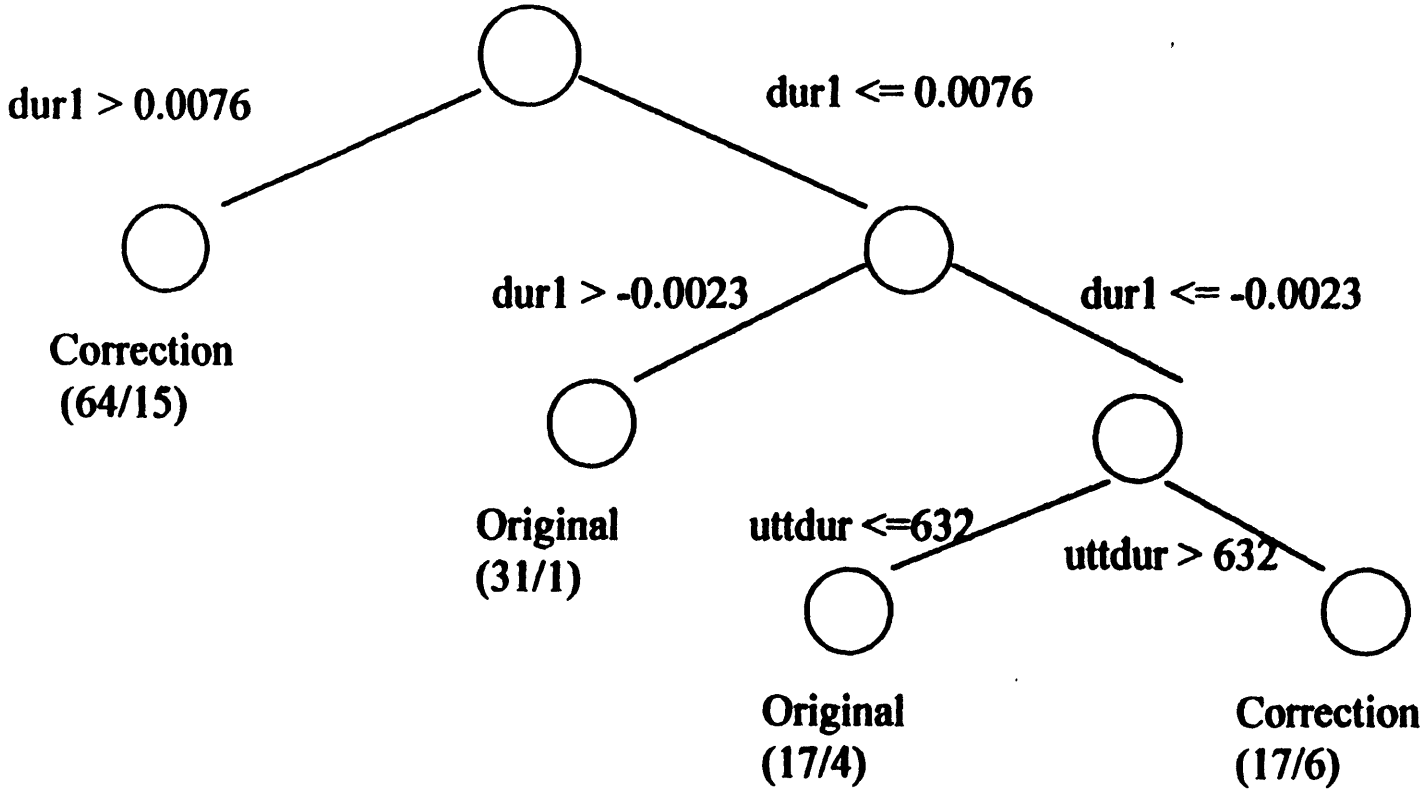
[3]In more general terms, the acoustic speaking rate measure uses the contrast between vowels, with fairly sinusoidal waveforms, and consonants, characterized by noisy waveforms to determine the number of phonemes.

In addition to these duration measures, pause measures also played an important, though secondary role in the general recognition of corrections. Average pause duration, or length per pause, greater than 70 milliseconds led to a classification of an utterance as a repeat correction in several classifiers. Likewise, the **pausevlen** measure of the proportion of an utterance composed of silence, also led to classification as a repeat correction when pause constituted more than 6% of the utterance (66% accuracy). These measures became important in decision trees in which only acoustic or syllable rate measure were available.

Example Tree



Misrecognition Classifier



5.3.3 Recognizing Corrections: Misrecognitions Only

We next consider separating original inputs from corrections of misrecognition errors only. We observed, in the course of acoustic analysis, that in pitch slope and accent measures, corrections of misrecognition errors patterned differently from corrections of other error types. The best classifiers for these corrections achieved 77.2% accuracy, or a 22.8% error rate. The training set error rate for the same trees averaged between 10% and 15%. These classifiers relied upon durational and pitch measures, specifically measures of absolute and normalized duration, pitch minimum, and pitch slope. In all cases, the first split in the tree was based on some measure of normalized duration. The use of pitch information improved the overall classification accuracy approximately 6%. The normalized duration measure used for the best result was `lenvexpvowel`, as discussed in the preceding section, normalizing total utterance duration based on the an average expected original input length for given text. [4] Using per-subject normalized syllable rate and `mrate` acoustic measures of speaking rate degrade performance to approximately 65%.

In all cases, the use of an absolute measure of pitch minimum and pitch slope measures, including steepest fall and sum of slopes. improved classifier performance over duration measures alone. The importance of these pitch measures in the classification of misrecognition errors contrasts strongly with their general negative impact on classification of other correction types. An example tree is shown below:

5.3.4 Recognizing Original-Repeat Pairs

Intuitively, an important factor in recognizing that a correction is taking place is the juxtaposition of the original input to the repeat correction. For instance, the contrastive use of accent observed in corrections of some misrecognitions involved using a different pitch accent type or increasing the amplitude of the existing pitch accent. Other obvious contrasts arise in durational and pitch change. Base speaking rate and base speaking pitch vary significantly across individuals. For instance, in the case of pitch, a low pitch minimum for a female speaker would look like a high pitch maximum for a male speaker. Some of the contrasts can be compensated for by per-subject normalization as we showed earlier in the case of pitch, but substantial variability remains. These contrasts can best be captured by treating the original-repeat identification process as the identification of the *pair*, as opposed to identifying the original or correction in isolation. One can view this approach as parallel to the identification of self-repairs, where one identifies the transition from the reparandum to the repair. [Nakatani and Hirschberg, 1994], [Shriberg *et al.*, 1997]

In this section we make a preliminary test of the possibility of using this pairwise information to identify original (position 1)- repeat (position 2) pairs, in contrast to

[4]The average original utterance duration for utterance text was computed from instances in the aligned data set. The number of instances available was thus relatively small, ranging from 1-20 instances per utterance text. Alternative estimates of original utterance duration can be computed from phoneme or word duration models, adjusted for position in word and utterance as in [Chung, 1997] or from a larger labeled corpus.

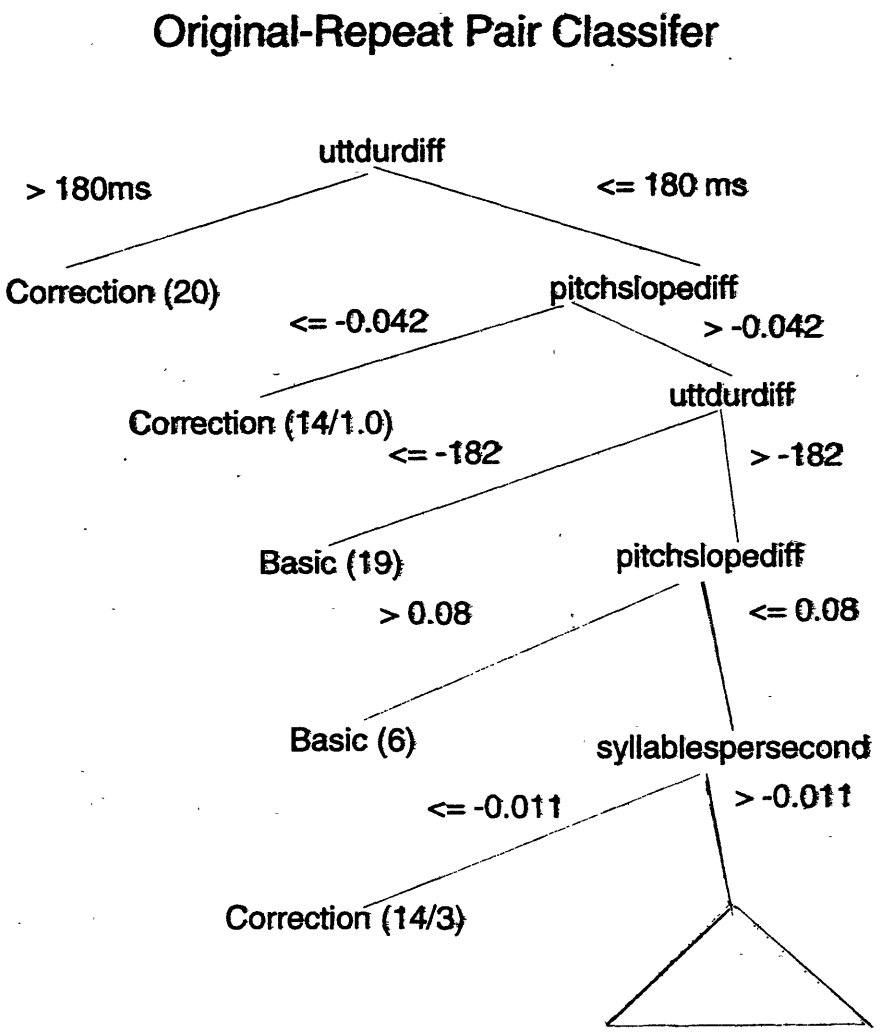


Figure 5-1: Original-Repeat Pairwise Classifier: Classified as Correction or Basic

"flipped" pairs, where the repeat is treated as occurring in position 1 or preceding the original (position 2). Here we substituted the difference of the features of the utterance in position 1 minus the features of the utterance in position 2 of the pair for the feature values themselves. We then performed a similar decision tree training and testing process to that in the previous experiments. For this admittedly simplified test, we achieved approximately 80% accuracy, over all error correction types. Increases in duration measures from original (position 1) to repeat (position 2) are the most important features and constitute the first split in the decision trees. An example appears below. 5-1

While this test of pair-based classification is overly simple, it suggests how to improve correction identification by using the contrasts between specific members of a hypothesized original-repeat correction pair. A more extensive test suite including examples of observed lexically matched *original-original* pairs to test against original-repeat or "flipped" pairs.

5.4 Discussion

We have seen that repeat corrections can be distinguished from original inputs at an accuracy of between 65-77%, depending on the type of error being corrected and the type of information available. While durational information plays the most important role in identifying all error correction types, the secondary information that yields improved results depends upon the type of error being corrected. When corrections of rejection errors are included in the classification task, pause measures, both average and proportion of utterance, play a role. When only corrections of misrecognitions are tested, pitch-related measures, such as slope and pitch minimum play important, though secondary roles. We also find that it is necessary to require a minimum number of nodes, usually between 8 and 20, for a sensible test in the course of building the decision tree. This constraint raises the error rate on training data from about 4% on fine-grained tests, requiring only 2 nodes per test, to the observed 10-17%. However, this constraint avoids overfitting which otherwise arises and can reduce accuracy on unseen data to chance levels, by limiting splits to subtrees classifying larger numbers of instances. While there are strong statistical trends to increased duration and pause, for example, the combination of some uncertainty, per-subject variability, and not strictly rectangular decision regions requires the implementation of constraints to avoid overfitting.

To put these results in a clearer perspective, let us consider an experiment reported by [Rudnicky and Hauptmann, 1990]. In this experiment, original input utterances and correction utterances of misrecognition errors from test interactions with a command-and-control task testbed with voice input and text output, were presented to experimental volunteer subjects who were asked to identify each utterance as a correction or not. These subjects correctly classified these utterances in 79.4% of cases, establishing a baseline for human performance on this task. Thus the range of accuracy reported for the decision trees, between 65-77%, represents a significant step toward human levels of performance, even where pure acoustic-prosodic measures were used. However, these classifiers are far from perfect, and since the task is non-trivial even for human users, we must consider how to make use of this limited but useful information. The most conservative approach to using this information would be to shift the style of interaction with the system. In other words, when the system suspects a correction is taking place, it can explicitly prompt the user for confirmation of its recognized utterance string. In general, experienced users prefer a less directive style, but [Oviatt *et al.*, 1994] have found that many users prefer a more restricted style, especially when recognizer error rates are high or they are unfamiliar with the system. Another way of using this information about the possibility of a correction taking place is to deploy a recognizer specialized for correction utterances to attempt to revise or verify the recognized utterance. One could, for instance, compare confidence scores for the different recognizer results, and select the one with higher confidence. A detailed analysis of the differences between observed durations and phonological realizations in correction utterances and those in a typical speech recognizer model will be presented in the next chapter.

Since we expected recognition of corrections to be easier with more contextual

information, specifically when the corrections can be directly contrasted with the corresponding original input in context, paralleling the recognition of self-repairs based on comparing the reparandum and the repair, we performed a preliminary investigation of pairwise original-repeat identification. Even in this simplified experiment we found improved rates of recognition of original-repeat pairs, suggesting that pairwise methods for identifying corrections may be a profitable avenue to explore.

Chapter 6

Implications for Speech Recognizer Design

6.1 Motivation

We observed in the initial discussion of the collected data that there was a large disparity between the probability of a recognition error occurring after a correct recognition and the probability of an error following a failed recognition, This contrast leads to the phenomenon of error "spirals", in which multiple successive recognition errors arise. These error spirals are particularly frustrating for users; user evaluations of spoken language systems have been shown to be closely tied to the ease or difficulty of correcting recognition errors. In a study of correction strategies in which extended error spirals were simulated [Oviatt *et al.*, 1996], the designers of the study found it necessary to scale back the maximum spiral depth (number of successive failures) to six, from an original depth of ten, when even pilot subjects became so frustrated that they refused to complete the tasks.

USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8 since yesterday.
USER: Give me the price for AT&T.
SYSTEM: Data General is at 63 1/4, up 1/3.
USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8.
USER: Give me the price for American Telephone and Telegraph.
SYSTEM: Excuse me?

In the SpeechActs field trial, error spirals proved to be a common problem for users. One subject encountered a sequence of 15 consecutive recognition failures, to each of which he responded with another attempt at the same utterance, showing remarkable persistence. In fact more errors occurred within the spirals than in first repeat correction position. Clearly, the adaptations that users employ to correct recognition errors in many cases seem to yield the opposite result.

As we observed in the previous chapter, these adaptations can be used to identify the corrective force of these utterances, which could not be recognized solely by lexical marking or repetition of lexical content. Clearly these changes provide useful and

necessary information to properly interpret the user's intent in uttering the sentence. We argue that it is, in fact undesirable to train users to avoid these adaptations; it is also difficult to do so. Users are often opaque to system directions; a classic example is the oft-reported difficulty of eliciting a simple "yes" or "no" response from a user, even when the user is explicitly prompted to do so. However, just as we note the utility of these cues for interpreting the corrective force of the utterance, we must recognize the severe negative impact that they have on speech recognizer performance. We will demonstrate that the systematic adaptations of users in the face of recognition errors that have been detailed in the preceding chapters have specific implications for the design of speech recognizers that will be more robust to the types of changes characteristic of correction utterances.

6.2 Duration-related Changes

In the analysis and classifier chapters, we noted three classes of systematic changes between original input and repeat correction utterances. There were (1) significant increases in duration, (2) increases in pause measures, and (3) significant decreases in utterance-wide normalized pitch minimum. Most contemporary speech recognizers strip out and normalize for changes in pitch and amplitude; thus pitch and amplitude effects are less likely to have a direct impact on recognizer performance, though pitch features do prove useful in identifying correction utterances. Thus, in this discussion, we will focus on effects of duration and pause changes that can impact recognition accuracy by causing the actual pronunciation of correction utterances to diverge from the speaking models underlying the recognizer.

6.2.1 Phonetic and Phonological Changes

One of the basic components of a speech recognizer is a lexicon, mapping from an underlying word or letter sequence to one or more possible pronunciations. In conjunction with a grammar, this lexicon constrains possible word sequences to those that the recognizer can identify as legal utterances. There is a constant tension in speech recognizer design between creating the most tightly constrained language model to improve recognition accuracy of those utterances covered by the model and creating a broader-coverage language model to allow a wider range of utterances to be accepted but increasing the perplexity of the model and the possibility of misrecognitions.

In addition to examining the suprasegmental features of pitch, duration, pause, and amplitude discussed in preceding chapters, we also examined phonological contrasts between original inputs and repeat corrections. We found that more than a third of the original-repeat pairs exhibited some form of phonological contrast, to various extents. For subsequent discussions we will divide these changes into two classes: one class deals with changes along what may be called a conversational-to-clear-speech continuum, as discussed in [Oviatt *et al.*, 1996], and another class deals with syllabic or phonemic insertions.

In the first class of phonological changes, we found contrasts between the classic

dictionary or citation form of pronunciation of the utterance, usually in the repeat correction, and a reduced, casual, or conversational articulation most often in the original input. Some examples illustrate these contrasts. Consider, for instance, the utterance "Switch to calendar." The preposition 'to' is a common function word, and this class of words is usually unstressed or destressed and surfaces with a reduced vowel as 'tschwa', even though the citation form is 'too'. Similar reductions are found with a variety of function words, e.g. 'the' which usually appears as 'th schwa' or 'a' as 'schwa'. Throughout the data set of original-repeat pairs we find more than 20 instances of a shift from reduced vowels, surfacing as 'schwa's in the original input utterances, to unreduced and occasionally stressed vowels in the repeat correction utterances.

These reduced-unreduced contrasts are not limited to vowel instances; a similar phenomenon takes place with released and aspirated consonants. For instance, 't' in the word 'twenty' can fall anywhere along a continuum from essentially elided (unreleased) 'tweny' to flapped 'twendy' to the released and aspirated of citation form 'twenty'. These contrasts are also frequent in SpeechActs data, as in 'nex' in an original input becoming 'next' in a repeat correction, or the frequent elision of the 'd' in goodbye, most often in original inputs.

In the contrasts discussed above we observed a shift from a reduced, conversational form in the original input to an unreduced, clear speech form in the repeat correction utterance. In this section, we discuss contrasts involving a shift from either citation or reduced form to an instance of syllabic or phonemic insertion. These instances arise in cases of extreme lengthening often accompanied by oscillation in pitch, similar to a calling pitch contour [Nakatani and Hirschberg, 1994]. A typical example would be the word 'goodbye' that surfaces as 'goodba-aye'. Approximately 24 instances of this type of insertion occurred in the data between original inputs and repeat corrections.

6.2.2 Durational Modeling

The conversational-to-clear speech contrasts and insertion processes discussed above are all phonetic and phonological changes which derive from a slower, more deliberate speaking style. In this section we will discuss how the increases in duration and pause described in the acoustic analysis chapter play out in terms of differences between observed utterances durations and speech recognizer model mean durations. We will demonstrate large, systematic differences between observed and predicted durations. This disparity is a cause for concern in speech recognition. In scoring a recognition hypothesis , two measures play significant roles: the score of the frame feature vector as a match to the model feature vector of the speech segment, and a timing score penalty assessed on phonemes that are too long or too short in the Viterbi decoding stage. In other words, recognition hypotheses will be penalized based on the amount the observed duration exceeds the expected duration. We will show that such a mismatch arises for a majority of the words in correction utterances and greater than two-thirds of the words in final position in correction utterances, where correction and phrase-final lengthening effects combine.

We obtained mean durations and standard deviations for a variety of phonemes

[Chung, 1997]. For each word in the SpeechActs data set we computed a mean and long, *mean* + *standarddeviation* measures of predicted duration by summing the corresponding means or long durations for each phoneme in the word. These mean and long word duration measures were then compared to the observed word durations in each of the original input and repeat correction utterances in the data set. We then reported the number of words exceeding the predicted mean and long durations and the average difference between the observed and predicted durations. [1] In addition, we computed the measures separately for words in utterance-final position, where, due to phrase final lengthening and the predominance of content words, we expected durational changes to be at their clearest. We present the durational shifts in original and repeat utterance as shifts from model duration in terms of number of standard deviations from the mean.

The first figure below presents histograms for all words, with the originals in dark grey and the corrections in light grey. Each point on the x-axis is one-half standard deviation, ranging from 1 standard deviation below the mean to 5 standard deviations above the mean. The first figure corresponds to utterances for all correction types. Note, there are very few instances of words less than the mean and also none less than a standard deviation below the mean. There is a large peak for the durations just slightly above the mean, corresponding to values between the mean and one-fourth standard deviation above the mean. The remainder of the words, approximately one-half for all correction types, exceed the mean by at least a standard deviation. The mean value for words in original inputs is 1.0987 standard deviations above the model mean; the median is at 0.8678. In contrast, for correction utterances, the observed mean rises to 1.353 standard deviations above the mean; with the median value at 1.0750. This shift represents a significant increase in durations. ($t = 3.6$, $df = 1398$, $p < 0.0005$).

The above figures raise the following question: what is the source of this difference from the model durations? It is clearly exacerbated for the repeat corrections, but it is also very much present for words in original inputs as well. Is it simply that the TIMIT durations are a terrible match for conversational, SpeechActs utterances? Or is there a more general explanation for the problem?

To answer these questions, we further divide the word duration data into two new groups: words in last position in an utterance and all other words. We saw before in the analysis of pitch contour the need to separate out utterance-final contours from other pitch accenting in the utterance in order to properly understand pitch phenomena. In addition, phonology argues that phrase- and utterance- final regions undergo a process referred to as phrase-final lengthening, which increases durations in words preceding phrase boundaries. In fact, one of the goals of [Chung, 1997] was to identify meta-features, such as phrase finality which might change the expected duration of phonemes.

First we look at histograms contrasting shifts from the mean duration for original inputs and repeat corrections for words in non-final position. Graphs for words from

[1] The durations of a small number of words with initial unvoiced stops may have been affected by the conservative approach to marking initial closure, used for pause scoring.

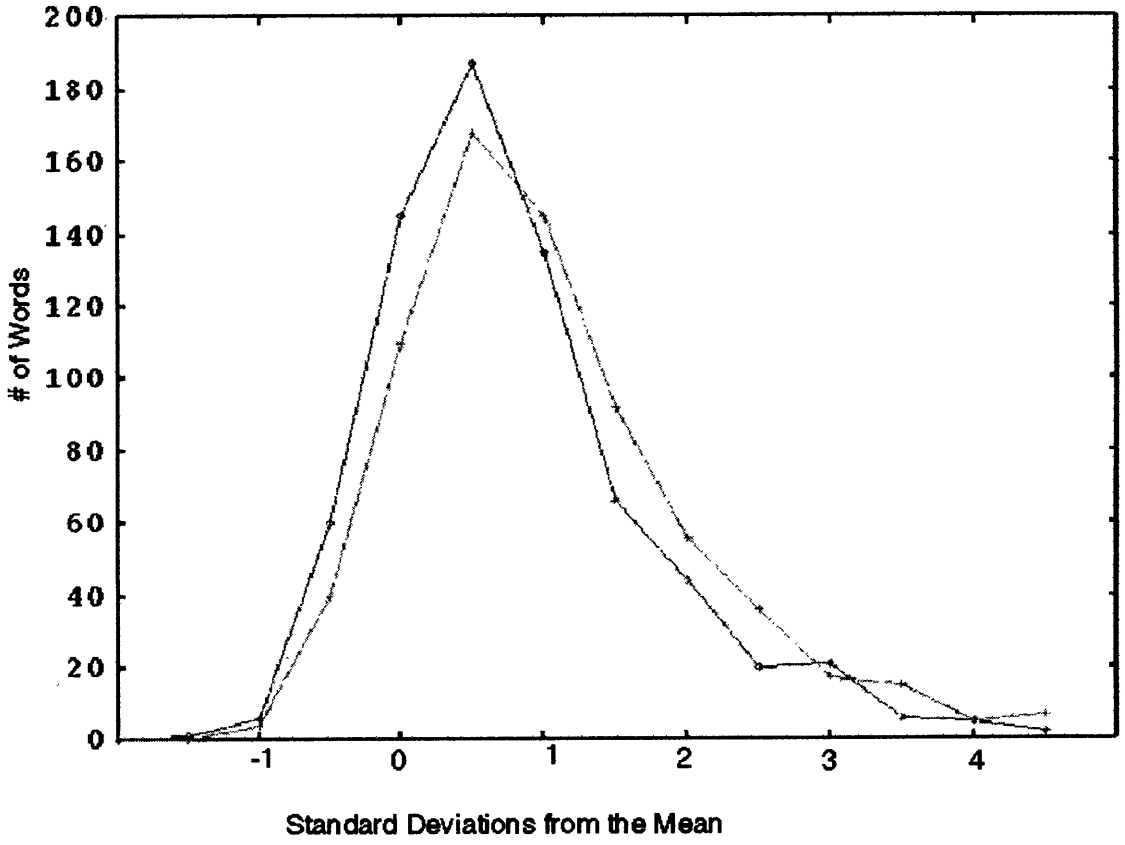


Figure 6-1: Overlapping Histograms: All Correction Types: Original (dark grey) and Correction (light grey): Histogram of Word Duration Shifts from the Mean, in Standard Deviations.

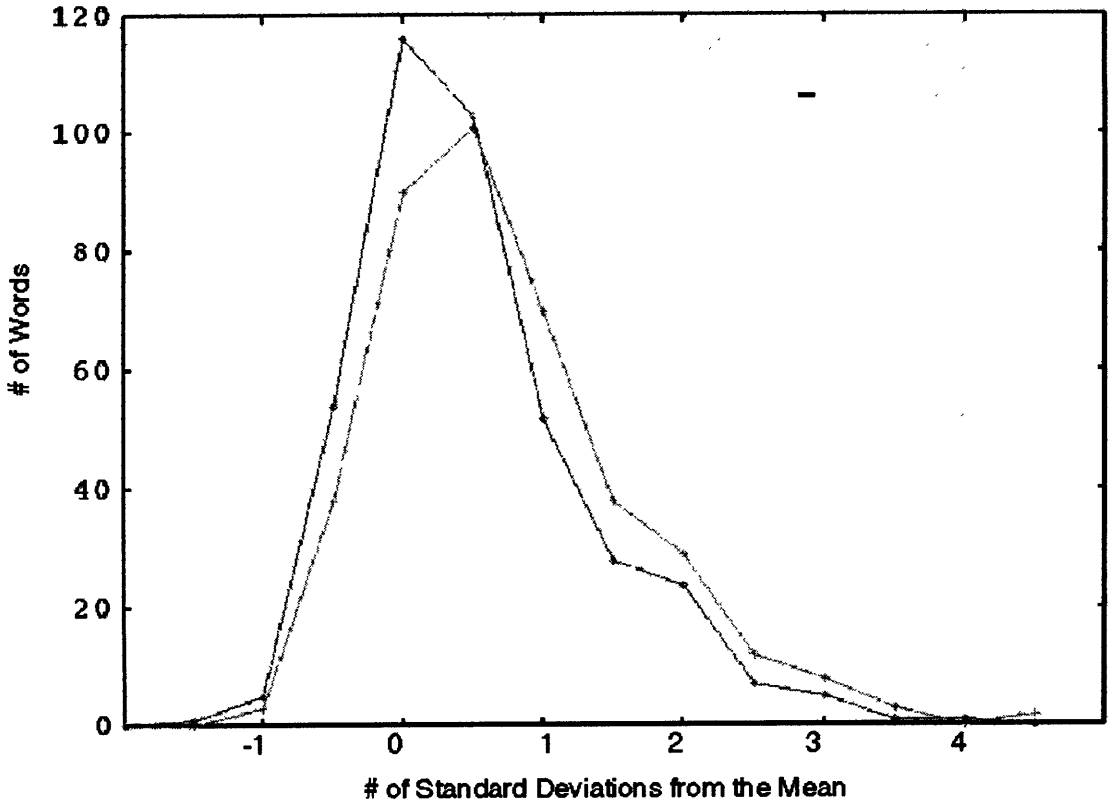


Figure 6-2: Overlapping Histograms: All Correction Types: Non-Final Words Original (dark grey) vs Corrections (light grey) Durations Distribution

all correction types (Figure 6-2) and corrections of misrecognitions only (Figure 6-3) are shown below. These figures contrast strongly with the distributions for all words. Instead, the distribution has a single large peak and two fairly narrow tails. In fact, these durations appear to be in closer agreement with the model, aside from having a slightly higher average duration with most durations falling between the mean and one-quarter of a standard deviation above the mean. The observed means for original inputs in non-final position are 0.7894 and 0.5520, and medians at 0.6404 and 0.4348, for all correction types and corrections of misrecognitions only, respectively, closer to the expected duration model. Secondly, we should note the difference between the distribution for words in original inputs and for words in repeat corrections, for non-final positions. The positions of the highest and second highest peaks reverse, placing the largest peak for correction utterances at approximately one-half standard deviation above the mean. Quantitatively the contrast between original and repeat inputs is even more apparent. The means rise from 0.7894 to 1.0556 for corrections of all types, and from 0.5520 to 0.7565 for corrections of misrecognition errors. These increases reach significance for corrections of all types (T-test: two-tailed, t= 3.3, df = 792, $p < 0.005$), and approach significance for corrections of misrecognition errors (T-test, two-tailed: t = 1.65, df = 204, p = 0.0518).

Now we examine only those words in utterance-final position, again displaying overlapping histograms for the distribution of durations for original inputs and re-

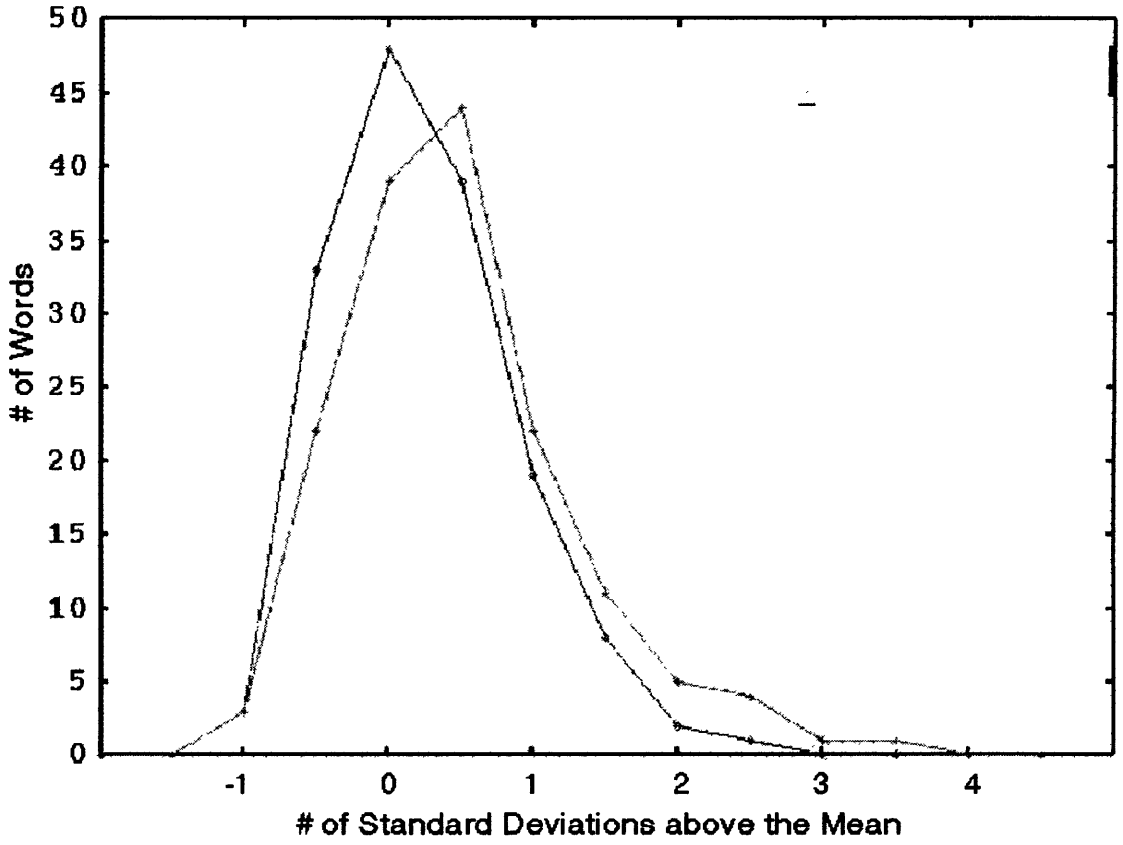


Figure 6-3: Overlapping Histograms: Corrections of Misrecogntions: Non-Final Words Original (dark grey) vs Corrections (light grey) Durations Distribution

peat corrections. Again we observe strong contrasts with the preceding figures. As suggested by phonological theory and [Chung, 1997]'s analysis, there is a significant increase in duration of words in final position relative to a general mean duration. Instead of a large peak less than one-quarter of a standard deviation above the mean, the largest peak for original inputs has shifted to between one-half and three-quarters of a standard deviation above the mean, depending on the error type. Not only is there a shift for the original inputs, but the words drawn from the repeat corrections shift even further.

Shifting to a more quantitative analysis, we find that the mean value for words in final position in original utterances is double the value for words in non-final positions. A similar relationship holds for repeat corrections, with corrections of misrecognition errors experiencing a greater increase.

Correction Type	Repeat?	Non-final	Final
All Types	Original	0.7894	1.5039
All Types	Repeat	1.0556	1.7446
Misrecognitions	Original	0.5520	1.1358
Misrecognitions	Repeat	0.7565	1.514

All of these contrasts between words in final and non-final positions are highly significant. (T-test: two-tailed, $p < 0.001$) These two groups should thus be viewed as coming from different distributions. The largest portion of the durational contrast between original inputs and repeat corrections arises from further increases in duration to the already lengthened words in phrase-final position.

The first graph below (Figure 6-4) illustrates the distributions for utterance-final word durations for corrections of all error types. The second graph (Figure 6-5) illustrates the analogous distribution for corrections of misrecognitions errors alone. We observe not only an overall rightward shift in the distributions for all repeat corrections in contrast to original inputs, but also a difference between the two groups of corrections. While the highest peak for corrections of all types decreases in amplitude with more 66% of words exceeding the mean by more than one standard deviation, the change for corrections of misrecognition errors is even more dramatic. The position of the highest peak actually increases by one-quarter of a standard deviation moving the distribution closer to a normal distribution (kurtosis = 3.0883, skewness = 0.4759, the lowest such measures for all distributions), centered now at one standard deviation above the expected mean. Both of these increases from original to repeat correction are shown to be significant. (T-test: two-tailed, $t = 2.07$, $df = 604$, $p < 0.02$ for corrections of all types and $t = 2.73$, $df = 174$, $p < 0.005$ for corrections of misrecognitions only).

This more detailed analysis of distribution of word durations in original inputs and repeat corrections allows us to construct a more unified picture of durational change. Basic duration models hold fairly well for pre-final words in original inputs, and show an increase to between one-fourth and one-half standard deviation above the mean in repeat corrections. In contrast, utterance-final words are very poorly described by these models. In all utterances the final words are subject to the effects of phrase-final lengthening, causing them to deviate from the models which suffice for other positions

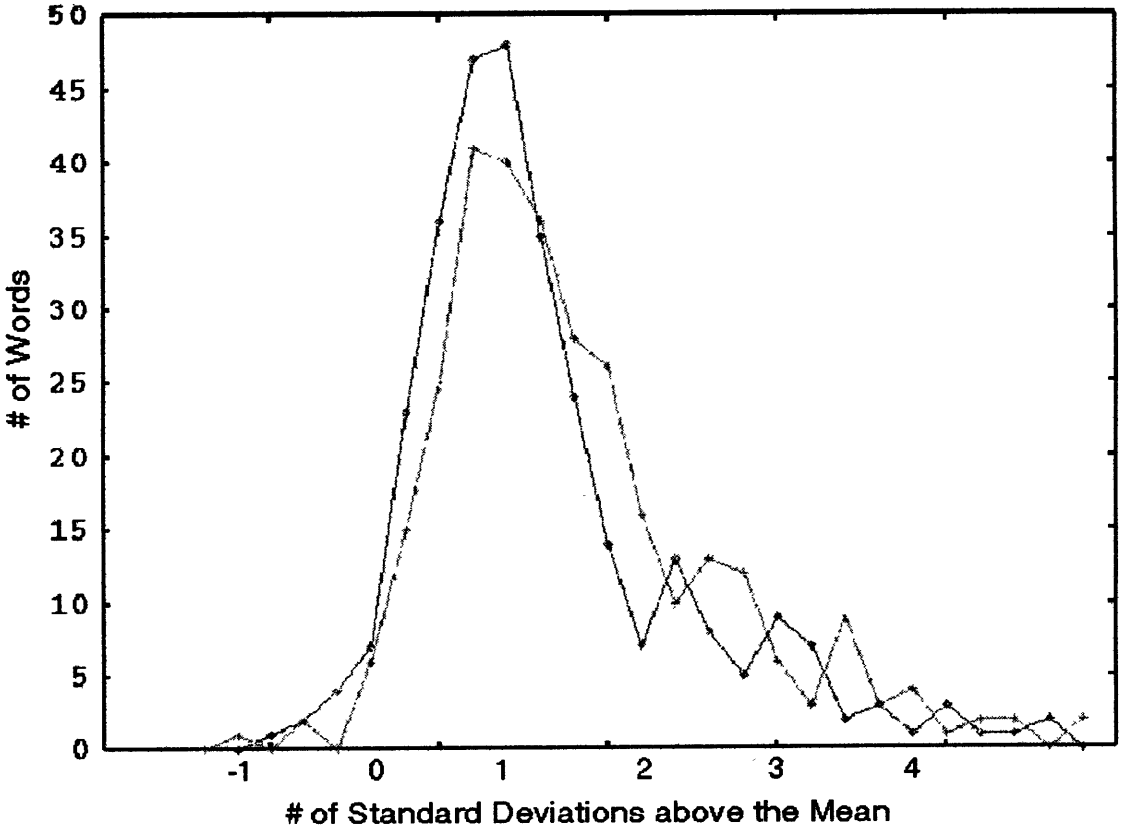


Figure 6-4: Overlapping Histograms: All Correction Types: Final Words Only Original (dark grey) vs. Correction (light grey) Duration Distribution

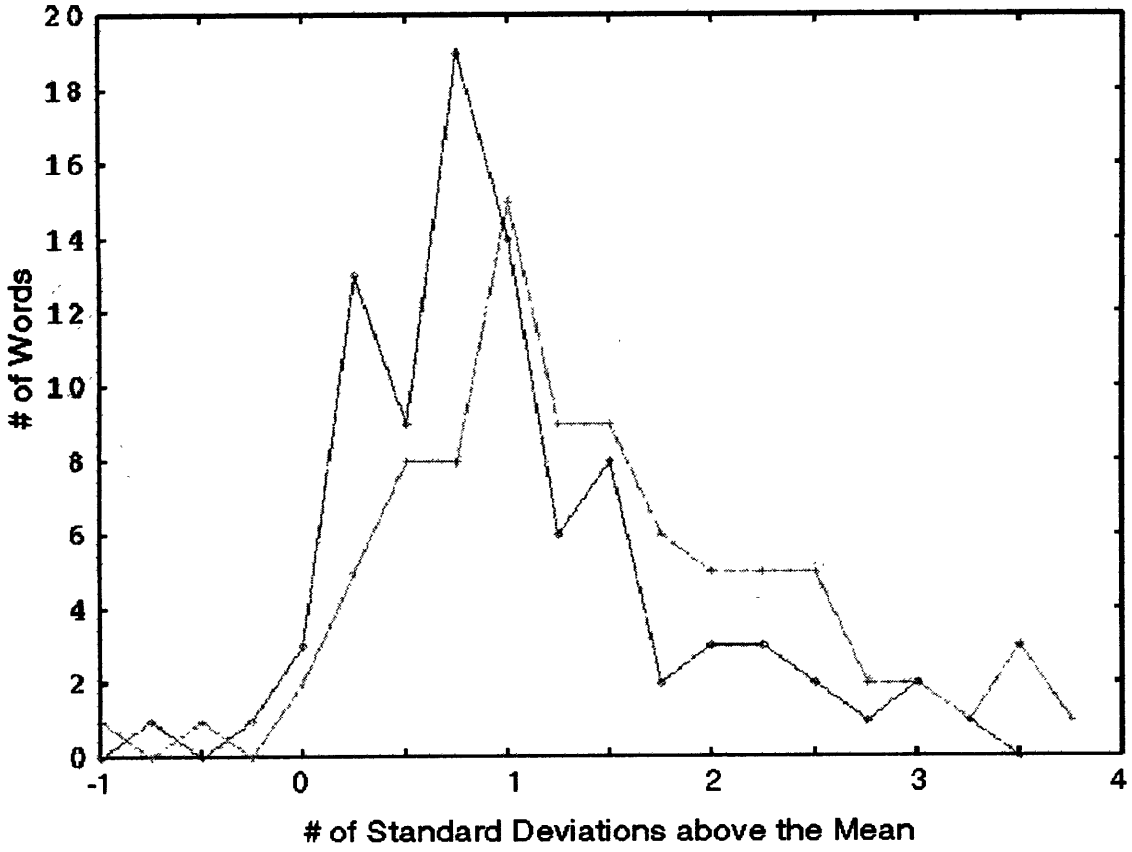


Figure 6-5: Overlapping Histograms: Corrections of Misrecognitions: Final Words Only Original (dark grey) vs. Correction (light grey) Duration Distribution

within the utterance. In addition, the effects of corrective adaptations, in turn, interact with and are amplified by the effects of phrase final lengthening. These combined effects cause words in utterance- final position of repeat corrections to deviate most dramatically from models of duration that do not take these effects into account. We see that these changes are most evident in corrections of misrecognition errors where a contrast with basic speaking style is most needed to inform the listener of corrective intent, in the absence of cues available for corrections of rejection errors where the system itself is aware of the recognition failure. Finally, the dramatic changes to utterance-final duration under the dual effects of phrase-final lengthening and corrective adaptation indicate the need for a durational model for speech recognition that can take this meta-information, such as position in utterance and discourse function, into account and further provide a starting point for the implementation of such a model.

6.3 Summary

The changes in speech and pause duration that we observed in acoustic analysis reflect not only a contrast between original inputs and repeat corrections but a shift away from the models underlying a speech recognizer. Phonological changes from reduced to citation form, following a conversational- to-clear speech continuum and further from citation to hyper-clear form through vowel and syllabic insertion move counter to the painstakingly modeled co-articulation effects of conversational speech and to a basic lexical model. In addition we observe a very skewed distribution of word durations, that in the change from original to repeat correction moves even further from the basic predicted phoneme durations.

Chapter 7

Conclusions and Future Work

Speech recognition error are a persistent and unavoidable problem for contemporary spoken language systems, often resulting in a 25% utterance error rate, causing one utterance in four to result in a recognition failure. These frequent errors are a source of frustration for users, creating a significant hurdle for the widespread acceptance of this technology. A vivid example of the problem caused by recognition errors, combined with a decrease in recognition accuracy for correction utterances, is the phenomenon known as an error spiral, seen below, in which the user repeatedly encounters errors while attempting the same command.

USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8 since yesterday.
USER: Give me the price for AT&T.
SYSTEM: Data General is at 63 1/4, up 1/3.
USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8.
USER: Give me the price for American Telephone and Telegraph.
SYSTEM: Excuse me?

This problem motivated the work in this thesis to identify the reason for this increase in recognition errors in correction situations and to attempt to defuse error spirals by identifying the presence of a correction, even when the system believes that successful recognition has occurred. Realizing that accurate recognition results may not be available and that, in any case, the interaction style and often limited vocabulary chosen by users of spoken language systems leads to many *non-* corrective repetitions, we have concentrated on the use of acoustic-prosodic features, measures of duration, pitch and loudness what are generally ignored in current spoken language systems, in correction utterances.

We have analyzed repeat corrections in contrast to original inputs in the context of a speech-only conversational telephone interface, SpeechActs developed at Sun Microsystems Laboratories. By providing a detailed characterization of acoustic-prosodic changes in correction utterances relative to original inputs in human-computer dialog, this thesis contributes to natural language processing and spoken language understanding. We present a treatment of systematic acoustic variability in speech recognizer input as a source of new information, to interpret the speaker's corrective intent, rather than simply as noise to be normalized or a bad habit that the

user should mend. We demonstrate the application of a machine-learning technique, decision trees and achieve accuracy rates close to human levels of performance for corrections of misrecognition errors, using acoustic-prosodic information to identify spoken corrections. This process is simple and local and depends neither on perfect transcription of the recognition string or complex reasoning based on the full conversation. We further extend the conventional analysis of speaking styles beyond a read versus conversational contrast to extreme hyper-clear speech, describing divergence from phonological and durational models for words in hyper-clear speech.

By analyzing acoustic-prosodic features of more than 300 pairs of original inputs and their repeat corrections, matched on subject and lexical content, we find significant differences between corrections and other inputs. Specifically, we find significant differences in durational measures, such as total utterance duration, total speech duration, and total pause duration. We also find an overall drop in pitch minimum along with an increase in the number of falling final intonation contours. All of these changes can be viewed as shifting from a more conversational or casual style to a more precise, careful, and clear speaking style. However, there are differences between clear corrective speech to computers and other clear speech environments. Increases in duration and pausing are the most consistent changes; while for some populations, there may be increases in loudness (the hard-of-hearing) or increases in pitch range (children in motherese[1] or even decreases in pitch as we see with computers. This study adds to our understanding of clear speech adaptations.

For one type of corrections, corrections of misrecognition errors, we also find an increase in pitch variability and accenting within the utterance, contrasting both with original utterances and with corrections of rejection errors. We trace these changes to a contrastive use of pitch accent marking the word or phrase where the computer's recognition error substituted an incorrect word. This adaptation fits well with current theory about marking new information with pitch accent, but is a change essentially orthogonal to the general conversational versus clear speech contrast of other acoustic adaptations seen in corrections.

We next demonstrate that these significant acoustic-prosodic differences between original inputs and repeat corrections can be used to develop a classifier to identify correction utterances. Using features derived from absolute and normalized forms of the contrasting measures, we train decision tree classifiers to identify corrections. We exploit the technique's robustness to irrelevant attributes and easy interpretation. These classifiers achieve accuracy rates between 65-77%, depending of the type of correction and the amount of information available. The best results are for corrections of substitution misrecognition errors, when a full text transcription is available; however, the methods still achieve 65% accuracy, over a 50% chance baseline and relative to 79.4% human performance level, when given only acoustic information. It is particularly heartening that the best results are for corrections of misrecognition errors, since the system, in these cases, would otherwise believe that no misrecognition had

[1]Motherese is a common term for the speech of care-givers to very young children. A specific speaking style characterized by expanded pitch range, increased duration, and higher average has been observed across several languages, though predominantly for female speakers.

occurred.

Finally, we look at the implications of corrective adaptations for speech recognizer design. We find that corrections are often characterized by phonological changes that shift from a more casual style, using reduced vowels and consonants, to a more clear, citation style, where the fully articulated forms are used. These changes parallel the conversational-to-clear speech shift noted for durational measures. In some corrections, insertion of a vowel or syllable extends these shifts to a "hyper"-clear speaking style, in which more sounds occur than in the dictionary form. All of these changes diverge from the basic speech recognizer model in which conversational co-articulation effects are painstakingly modeled. Duration of corrections also diverges from a base speech recognizer duration model. Utterance-internal words in original utterances make a good match for a phoneme duration model based on TIMIT utterances, as in [Chung, 1997]. Sentence-internal words in repeat corrections show a systematic increase from this base model, and 75% of words in final position, where correction adaptations and phrase-final lengthening effects combine, are more than one standard deviation longer than the mean duration predicted by the model. Clearly, a speech recognizer model needs to adapt to discourse features such as corrective intent in addition to sentence-level effects such as phrase-final lengthening.

Clearly the analysis in this thesis has not completed our understanding of corrections, for recognition or modeling. The current data set, 300 original-repeat pairs of which two-thirds are corrections of rejection errors and one-third corrections of misrecognition errors, is still quite small. Several issues must be explored in greater depth. One important question involves variability across subjects. For example, while the overall trend was toward significant increases in duration, one subject consistently decreased the duration of his correction utterances relative to original inputs. In the area of pitch, while most subjects changed any rising contours in originals to falling contours in repeats, one subject displayed the opposite trend, switching all contours to final rises in corrections. Perhaps the most important feature of correction is the contrast with respect to original input. It may be most effective for data sets with more per-subject data to parameterize original-repeat correction modeling for each variable measure, such as pitch contour "change to fall" or "change to rise", and then set these parameters on a per-subject basis as the user becomes familiar with the system.

Another form of variability in corrections that we have not explored is whether the contrasts that we encountered for lexically matched corrections, increases duration and pause, decreased pitch minimum, hold similarly for corrections with different lexical content, as in "play message eight" (original) and "read number eight" (correction). The general trends of a shift to a more clear speaking style suggest that similar adaptations will occur, but the question should be answered explicitly.

Another natural extension to this work would examine the adaptations for corrections later in the error spiral. We hope that earlier detection will limit the number and depth of error spirals, but they will not be eliminated. [Oviatt *et al.*, 1998] presents some descriptive analysis of within-spiral corrections, indicating a decrease in variability. This analysis may suggest whether corrections deeper in the error spiral should be identified with respect to the preceding input attempt of with respect

to a general model of "original" and "repeat". These straightforward extensions will provide a more thorough understanding of spoken corrections and will expand our ability to detect corrections, improve our ability to detect discourse relations and improve error recovery in spoken language systems. In the following sections we consider future work that would lay the groundwork for a more complete system for speech recognition error recovery.

7.1 Pairwise Identification of Corrections

[Rudnicky and Hauptmann, 1990]'s experiment in which listeners identified utterances as corrective or not for utterances from human-computer dialogs established a baseline performance of 79.4% for identifying corrections in isolation. However, corrections in human-computer dialog are not made in isolation; in particular, the original, likely preceding, input is available for comparison in a spoken language system, without violating our goal of a local classifier. It seems that the inclusion of this narrow contextual information in the classification process can only help to improve the identification of correction utterances. Instead of comparing the utterance to be classified to a generic model of an original utterance, as is otherwise accomplished using normalized duration, speaking rate, and pitch measures, one could compare the possible correction to the hypothesized original of the pair. Such a direct comparison would likely improve over per-subject normalization of the measures. The simplified test of this pairwise approach reported earlier improved over the best isolated utterance classification results. A crucial requirement for such experiments would be a more extensive paired database, drawn, for instance, from all adjacent inputs in a human-computer dialog corpus. Such a data set would need to include examples of sequential, lexically matched *non*-corrective input, as well as lexically different correction and non-correction pairs, in addition to the type of data examined in this thesis.

7.2 Future Work: Building a System

7.2.1 Introduction

The core work of this thesis demonstrates that there are significant differences between original inputs and corrections. The differences in addition represent a divergence from underlying speech recognizer models, leading to the counterintuitive worsening of recognition accuracy when the user tries to speak more clearly. Furthermore, we can use these contrasts to train a classifier to distinguish corrections from other inputs. The natural practical extension of this research is to incorporate such a classifier in a spoken language system to detect and repair human-computer miscommunication. The decision tree classifier discussed in the thesis performs identifies that an error has occurred by the user's attempt to correct it. We propose two components to participate in the repair of the error: one, a context-sensitive speech recognizer that can compensate for speaker adaptations in correction utterances, and, two, a strategy

for error repair interactions between the system and the user that will help prevent error spirals and, in particular, the user's feeling that the system is not responsive to his efforts to make corrections.

7.2.2 Recognition: Practical Issues

The key first step in our model is to identify an utterance as being a correction. Once we make this determination, we can determine whether to deploy an adaptive recognizer, attempt to find a specific corrected word or phrase, and pursue a repair subdialog. Could we avoid this separate step by using the comparison of results from two recognizers, one adapted for corrections and one basic, to make this determination implicitly? It is probable that we can not. First, it is generally impractical to compare the scores of outputs from different speech recognizers; scores are ranked internally, relative to each other, not with respect to some absolute measure of quality. Secondly, such a design choice would limit the types of information available to the classification process. Specifically, speech recognizers normalize away the pitch variability that proved useful in identifying misrecognitions. Furthermore, we could not smoothly incorporate pairwise comparisons of utterance sequences as 'original-repeat', discussed earlier as a possible improvement over the difficult task of identifying corrections purely in isolation, into such a model, since the *difference* between utterances is important to the pairwise approach.

A decision tree classifier that uses acoustic-prosodic features, as described earlier, should be realizable within the framework, time, and resource constraints of a typical spoken language system. First, the time-consuming process of training and tuning the decision tree classifier can be done off-line. A developer must first collect a training corpus of at least 300 original-repeat correction pairs, either from a live system or through a "Wizard-of-Oz" style experiment. They must digitize, label, and transcribe the corpus. The acoustic analysis tools are widely available, including pitch tracking, silence detection, speaking rate measurement, and forced alignment, as is decision tree software. A confidence measure for the classifier would be useful; a weighted pessimistic estimate of the error rate for a given branch, based on the training set error and similar to that used for tree pruning, would be a good candidate. When we discuss error repair dialogs, we will see that a confidence measure would be useful in determining how aggressive the correction strategy should be.

In the context of the typical operation of a spoken language system, this approach should still be feasible. Many of the acoustic analyses are already performed as part of the speech recognition process, as is the case for silence detection and even pitch tracking in recognizers for tone languages like Chinese. The decision trees themselves are compact, with between 7 and 37 nodes, and thus relatively fast as well. Approaches that combine acoustic-prosodic measures and machine-learning classifiers in conjunction with speech recognition systems have been successfully deployed by [Ostendorf *et al.*, 1996](to distinguish read, conversational, and fast speech) and [Taylor *et al.*, 1996b](to identify speech act type and constrain recognizer domain), providing empirical evidence for the practicality of this approach.

Isolating the Correction

Once we have identified an utterance as corrective, we can attempt to more carefully isolate the site of the misrecognition. Identifying an utterance a correction of a misrecognition error is very useful, since it provides information to the system about how to interpret the utterance in accord with the user's intent, in part by creating uncertainty about the preceding action that necessitated the correction. However, it would be even more useful to be able to identify what part of the previous action was in error, to help facilitate the repair.

Corrections of misrecognitions can be further divided into groups based on how much and what part of the utterance was misrecognized. The amount of assistance the system can provide depends how much was wrong in the misrecognition. On one extreme of these subclasses are what have been referred to as "off-the-wall" errors [Oviatt *et al.*, 1996], in which the action of the system has no apparent connection to what the user actually said. An example of such an "off-the-wall" misrecognition would be the following:

USER SAID: Undo that.
SYSTEM HEARD: Goodbye.
SYSTEM SAID: Do you want to hang up?

Here the entire utterance is misrecognized, and there is little more information available about how to effect the repair than if the utterance has simply been rejected.

Other misrecognitions are more limited in scope, and can provide more assistance in a repair strategy. Consider, for example, the following sequence:

USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?
SYSTEM SAID: Fifteen US Dollars is 30 German marks.

Here only the number "fifty" is mistaken for "fifteen." Knowing what porting of the utterance is being corrected would allow a system to be much more helpful to the user in correcting the misrecognition. It could compare the recognition results for the original and correction, if available, in the suspect region to note whether the same error had occurred. It could also tailor prompts to elicit the problematic piece of information, and shift to a more directive style while giving feedback on the more stable part of the recognized utterance. In the above case, such help might take the following form: "SYSTEM SAID: Enter the amount of US dollars to convert to German marks."

As a first step toward this more helpful type of system, we could begin by isolating the point of misrecognition in utterances where the error is localized to a single word, as in the "fifty"/ "fifteen" confusion described above. We believe that these relatively discrete errors would be easiest to identify, since they are restricted to a single word rather than a phrase or several separated words in the utterance. These points of contrast are, moreover, the classic location for the use of contrastive pitch accent,

that should have the acoustic correlates of pitch movement, increases in duration, possible increases in amplitude, and possibly preceding pause ([Stifelman, 1993]).

We performed a simple experiment to test the plausibility of this approach. We selected a data set composed of all corrections for which there was a single isolatable error being corrected. We then labeled each word as basic or corrected. We used word-level variants of the features used in the earlier decision tree classifiers, augmented with word position information and adjacent word pitch, amplitude, and pause information. There were 28 isolatable correction words, and using a decision tree classifier, we identified 26 of the 28 correction words with 2 false alarms. The first split in the tree was on whether or not the word being classified was in utterance-final position. This approach thus seems quite promising, although the data set is far too small for the results to be viewed as other than preliminary.

7.2.3 Recovery: Improving Recognition

Having identified the utterance as corrective, we have performed the detection part of our task, and now we must turn to the process of recovering from the error. We will show that isolating the site of misrecognition as described above can play a key role in this activity. First, however, we look at improving recognition accuracy on corrections and then we will explore repair dialog strategies.

We observed in our data, as noted by other researchers [Shriberg *et al.*, 1992], that users experienced more recognition errors on corrections that other inputs. Our acoustic analyses of correction utterances demonstrated divergences from base recognizer models of conversational speech in both phonology and duration. Thus we propose a method to adapt the speech recognizer to perform better on the hyperarticulate speech that often characterizes corrections. One might consider simply augmenting the training data in a single speech recognizer with instances of correction utterances as well as original inputs. Unfortunately, while simple, this approach is impractical since it would simply over-generalize the recognizer, probably worsening overall recognition rates. The problem with such an approach is that it misses two important points: first, that this speaking style is not just a random variation, but occurs in specific speech acts, and further that the variation is systematic. There is a clear trend to increases in duration and clear speech phonology, nd these effects can extend throughout the utterance, rather than just as phoneme by phoneme changes. Furthermore, a two-part approach adapting both the phonological and durational models in concert is necessary. Adapting phonology alone would not improve recognition in those cases of large durational increase but no explicit lexical change; changes in phonology contribute to and amplify durational changes, but do not alone account for all variability. [Oviatt *et al.*, 1996] Adapting the durational model alone fails as well; durational models are based on phonemic identity and, in addition, common phonological changes are not uniformly distributed across phonemes or words, but, for instance, occur more frequently in function words. Only by modifying both the phonological model through the lexicon and the durational model can we hope to accurately model and recognize speech in correction utterances.

To compensate for phonological changes explicitly, we propose a set of phonologi-

cal rules that can be used to transform the lexicon for correction adaptations. These rules map from expected recognizer phonological realizations to those which actually surface in corrections. They can also incorporate some phoneme context or word type constraints, and could be augmented with a probability, indicating the likelihood of this form as opposed to the conversational. For example, a possible rule would be "schwa ==> full vowel, in function words" and would cover cases such as "t"' ==> "to". One would not want to just add the results of the transformations as alternate pronunciations to the lexicon for all utterances, not just corrections, because it would weaken the model by increasing perplexity in the general case and thereby worsen recognition accuracy. A similar technique of using phonological rules conditioned on a speaking style has been used by [Ostendorf *et al.*, 1996] to improve recognition accuracy in very fast or casual speech. We can derive these rules by generalizing from our observed phonological adaptations and also from general rules about co-articulation, since many of the correction-related changes reverse those of co-articulation.

To compensate for overall durational changes, we must modify the recognizer where durational information comes into play. As noted earlier, duration plays its main role in a speech recognizer at the Viterbi decoding stage. At this stage, the system attempts to select the recognition string with the best match score for the observed acoustic sequence. A recognition hypothesis is penalized whenever its constituent phonemes are shorter or longer than allowed by the durational model. There are thus two points at which a system designer can affect the use of durational information: explicitly in the phoneme duration model and in the penalty for duration mismatch.

Let us begin with the latter as it is the simplest. One could just decrease the amount of penalty assessed for phonemes that are too long. Naturally, the penalty should not be decreased too much; with no penalty for over-long phonemes, the recognizer could prefer a single phoneme with the best acoustic match to a sequence of phonemes. However, this approach by itself is not optimal, as it does not capture the systematic character and magnitude of durational increase. It also does compensate for the effects of position on duration, since the durational penalty is independent of location, using only the difference between the observed and expected durations.

A better approach would provide a more precise formulation of durational change. It should capture, at least, the two types of contrast observed: increase in duration from original to correction, and larger duration in final versus non-final word position. One could simply try to build a durational model explicitly from scratch from correction examples. However, sparseness of data could prove a problem, and, perhaps more importantly, such an approach could not take advantage of the systematic nature of these durational increases. [Chung, 1997] describes a hierarchical model of phoneme duration, based on word and sub-word units. Such a model can take into account effects of word, phrase, and utterance position on phonemes and has been shown to reduce durational variances and provide a more accurate model. A natural extension to this approach is to extend the hierarchy beyond the sentence level to incorporate effects of discourse structure and relation, such as corrections. As a first step toward such a model, we propose a duration model for corrections that modifies the hierarchy as a whole according to the systematic increases found in duration of ap-

proximately one-fourth standard deviation for non-final words and one-half for words in final position. A hierarchical model would also provide some predictive power for determining the relative changes for the different phonemes within the word, based, for instance, on phoneme type and syllable or affix status. This model can obviously be augmented or tuned by adjusting the global duration penalty and would provide the basis for a general context-adaptive recognizer.

7.2.4 Repair Interaction Strategies

Improving recognition accuracy on corrections to levels closer to those for original inputs through use of a specially adapted recognizer should decrease the frequency and length of error spirals. However, we would like to perform better and respond to the user's corrective intent. For instance, there are cases, such as out-of-vocabulary utterances, where even the best adaptive recognizer will still fail. Even when the correction is successfully classified and recognized, we may still wish to verify the user's input in order to block the possibility of further error spirals or if the correction requires that we retract a previous action.

We describe a group of repair interaction strategies that differ primarily in how aggressively they act to perform a repair and correspond to the system's confidence in its course of action. Specifically, we consider three levels of repair interaction: 1) confirmation, 2) elicitation, and 3) retraction. Confirmation is the policy of least intervention and least cost. In a confirmation strategy the system asks the user to verify the recognition result. An example from the currency exchange interaction would be:

USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?
SYSTEM SAID: Fifteen US Dollars is 30 German marks.
USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?
CLASSIFIED AS: Correction
SYSTEM SAID: Do you want the price of fifteen US Dollars in German marks?

Confirmation is a good strategy when the confidence for the classifier is low. If the user utterance was not a correction, it adds only one unnecessary query and response; if the user utterance was a correction, it provides additional confidence for the system in pursuing the repair at a higher level of intervention if necessary. The confirmation stage is clearly unnecessary if the correction itself caused a rejection error; there is no input to verify and no question that an error has occurred.

If there is higher confidence that the user has made a correction, either because of a better score from the classifier, the user said "no" at the confirmation stage, signaling that the assumed correction was itself misrecognized, or a rejection error has occurred, the system can move to the next level of intervention, elicitation. This level corresponds to a shift to a more directive interaction style or, in other words,

to greater system initiative. At this stage, the system can take a variety of different steps to elicit from the user enough information to perform the correct action. If the system has identified a misrecognition and isolated the site of the error, as described earlier, the system can prompt for that information explicitly.

USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?
SYSTEM SAID: Fifteen US Dollars is 30 German marks.
USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?
CLASSIFIED AS: Correction, site: fifteen
SYSTEM SAID: How many dollars do you want in marks? Speak or type.

Depending on the type of information being elicited, the system could ask the user to select an item from a list, user a telephone keypad to spell a name or type a digit, or even simply to spell the input aloud. These options try to exploit the difference in perplexity and confusability for different modalities in conjunction with a user tendency to mode-switch after a failed correction [Oviatt *et al.*, 1996] to improve chances of correctly recognizing the input. Other strategies for identifying the point of misrecognition are also available, including:

- Lexical comparison

  Comparison of the inputs before and after the suspected error, for instance, for a single word difference.

- Task constraints

  Asking for an source location for travel, when all other information is already known.

- Belief constraints

  Identifying a direct conflict between recognized input and internal belief.

Even when almost no information is available from the input, as in the case of a rejection error, the system can elicit information based on the context, either with a context-sensitive help message with example utterances or an explicit prompt for each field needed for a command in the current application context.

Finally, if the system has very high confidence that a correction has occurred, either from the classifier or confirmation requests, the system can pursue the most aggressive correction strategy, retraction. In this case the system must undo the previous action and any state change associated with it, unless that state is implicit in the correction.

USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?

SYSTEM SAID: Fifteen US Dollars is 30 German marks.
USER SAID: How much is fifty dollars there?
SYSTEM HEARD: How much is fifteen dollars there?
CLASSIFIED AS: Correction, site: fifteen
SYSTEM SAID: How many dollars do you want in marks? Speak or type.
SYSTEM BLOCKSfifteen.

Thus, depending on classifier confidence or explicit user confirmation, the system can intervene to varying degree to facilitate error repair and recovery. These strategies range from confirmation requests at the most tentative end of the spectrum to retraction of the previous action at the other. An intermediate level, elicitation, allows the system to exploit a wide range of information from acoustic and lexical cues to task and belief constraints to guide the user to an effective input in a system-initiative style, such as [Oviatt *et al.*, 1994] has found that users prefer in many circumstances when error rates are high or the user is uncertain. These strategies all pursue the goal of minimizing the depth of error spirals and providing feedback to the user that the system is cooperating to repair any errors that occur, while remaining sensitive to the possibility of correction classification error.

7.2.5 Correction Detection and Error Recovery

The work in this thesis combined with the future work outlined above forms the basis for a system for recovery from speech recognition errors. By providing a detailed characterization of the differences between corrections of recognition errors and original inputs, we allow the development of decision tree classifiers to detect correction utterances. The ability to detect corrections actually allows one to detect the presence of misrecognition errors by recognizing the user's response to a system error, where the system would otherwise be unaware of its mistake. Based on this knowledge that an error have occurred, the system can now shift to error recovery. It can utilize a speech recognizer that adapts to the speaking style employed in spoken corrections, rather than allowing the divergence between basic speech recognizer duration and phonological models and correction utterances to lead the system into a deepening error spiral. It can further use the technique described in future work for isolating the word being corrected to focus system help on eliciting that important piece of new information from the user. Even if a single error site cannot be identified, the recognition of a correction can move the system into a more structured interaction [Oviatt *et al.*, 1994] to guide the user through error recovery or signal that system help should be offered to the user. All these components can be brought together to build a system to facilitate error detection and recovery, focusing on one of users' most significant sources of dissatisfaction with spoken language systems, the difficulty of correcting errors.

Appendix A

C4.5

This section describes in more detail C4.5, the specific decision tree building algorithm used to train and test the classifiers discussed in this thesis. We discuss the choices of entropy measures used to select tree splits, pruning criteria, and construction of rules from trees. We also discuss the impact on feature design and decision tree parameterization of the specific form of the data for recognizing spoken corrections.

Let us begin with a statement of the basic decision tree building algorithm underlying any implementation. The algorithm begins with a set T of labelled training instances and a set of classes C1, C2, ..Cn to which these cases are to be assigned. The technique then recursively proceeds as follows:

- Case 1: All of the instances in T belong to a single class Ci.

  The decision tree for T is a leaf, and it labels instances as class Ci.

- Case 2: There are no cases in T.

  The decision tree for T is a leaf node. It labels instances as some class Ci, according to some heuristic.

- Case 3: The instances in T represent members of different classes.

  The decision tree is a branching subtree. The new branches should trend toward sets of instances that are more homogeneous. The goal is reached in the following fashion. A test, T1, is chosen that divides T into one or more mutually exclusive subsets, S1, S2,..,Sn. Test1 is a test on a single attribute or feature. For a discrete feature, this corresponds to producing one subset per possible attribute value assignment. For a continuous feature, this corresponds to selecting a dividing points in the range of values and creating subsets for V1 < F1 and V1 >= F1. The best test Test1 is the test producing the best score of improved homogeneity, with different algorithms using different scoring metrics. The new subtree is rooted at Test1, and the branches correspond to each possible outcome of the test. Each new subset of T, S1, S2,...,Sn associated with each outcome is treated recursively as T.

The approach described above is clearly a recursive divide-and-conquer algorithm. In addition the selection of a best test at each branching node implies a greedy algorithm. Since exploring all possible decision trees to find the most accurate and most compact, therefore the most predictive, is NP-complete, it makes sense to choose

a greedy approach and then try to select the best heuristic for choosing the dividing test at each stage.

A.1 Specifying the algorithm

Let us begin by stating the specific choices made in C4.5 for the underspecified portions of the algorithm above.

A.1.1 Case 2: T with no instances.

When T is empty, in other words, when a branch created by a test has no instances in the training set, the system must still predict a label for unseen instances that fall in this branch. Here the simple heuristic of labelling the branch with the most frequent class is chosen.

A.1.2 Finding a splitting point

For continuous-valued features, it is necessary to find a value at which to divide the feature values into two discrete subsets for branch tests. This step is particularly relevant for the classification of original and repeat correction utterances since the vast majority of our features are continuous-valued. We describe briefly how tests are proposed for such features. The system proposes a test splitting the continuous-valued attribute into two sections at the midpoints between each ordered pair of values for this attribute encountered in the training set. In other words, if there are n instances in T, there are at most n-1 possible positions for the split to be made. This approach is standard for most decision tree algorithms. C4.5 includes the restriction that the threshold value be the attribute value closest to, but not actually exceeding this midpoint, ensuring that the threshold value actually appears in the data set.

A.1.3 Measuring Homogeneity

Like many other decision tree implementations, C4.5 uses as information theoretic measur to assess the best test. These methods are generally more effective than those, such as those used in CSL1, where simple rubrics such as choosing the test that yields a pure branch are used. Such an approach would be extremely sensitive to noise. The information theoretic measure on which the splitting criterion is basd for many decision tree algorithms is entropy, a measure of the amount of information needed to encoded the class of a case. $\text{Entropy(s)} = -\sum_{j=1}^{k} \frac{\text{freq(Cj,S)}}{|S|} \log \frac{\text{freq(Cj,S)}}{|S|}$

ID3 introduced an extension of the entropy measure to determine the split that *improved* classification the most. This measure, called the *gain criterion* is the difference in entropy between the full set of instances at the root of the proposed subtree and the weighted sum of the entropies of the subsets produced by the test.

$\sum_{i=1}^{n} \frac{|Si|}{|S|}$ Entropy(Si) . This criterion, however, is biased toward tests with large numbers of outcomes. In particular, it would strongly prefer classification based on a unique identifier, since it produces subsets with 0 entropy , over any more predictive division. This bias is clearly undesirable. C4.5 compensates for this trend by normalizing with respect to the number of subdivisions, calculating an information measure for the split as, $-\sum_{i=1}^{n} \frac{|Si|}{|S|} \log \frac{|Si|}{|S|}$. This normalized gain ratio criterion is the measure to be maximized by the chosen splitting test. This measure has generally been found to be advantageous over the unnormalized form, and comparisons of splitting criteria [Mingers, 1989] indicate that it produces compact trees. In the case of our analyses, all but one of our features is either continuous or binary-valued; so the bias in the gain criterion is less important. Nevertheless, [Quinlan, 1988] notes that it is still experimentally advantageous to use the gain ration criterion in the case of binary tests.

A.2 Pruning

When the decision tree algorithm runs to completion, it will subdivide the data until no test yields an improvement or even until leaves classify only single nodes. Such behavior can lead to overfitting, degrading predictive power by making classifications based on few example instances. One step to limit such overfitting provided by C4.5 is pruning of the initially produced decision tree. While it might be more efficient to simply halt the tree division process before overfitting becomes a problem, as Breiman et al note, it is very difficult to design such criteria. Pruning , in general, simply replaces a subtree with a leaf or one branch of the subtree. The question is how to select a subtree to be pruned, preferably in such a way as to decrease the predicted error rate on new unseen cases. One could compute this predictive error rate by testing on some held out cases, as in "cost-complexity pruning" (Breiman et al) which selects nodes for pruning based on an MDL-like approach combining a measure of the complexity of the subtree and the training error rate, weighted based on the results on the unseen cases. An alternative method also using a held-out set is "reduced error pruning" (Quinlan 2), that directly computes the error rates on the unseen cases. However, in our constrained data situation, we use the method in C4.5 that computes a pessimistic estimate of the predictive error rate based on the observed error rate and measures of statistical confidence. This pessimistic error estimate treats the training error, X errors in N instances classified. as a probability of error, X/N. C4.5 then sets the predicted probability of error to an upper limit, for some confidence level, found for the (upper) confidence limit of the binomial distribution, Ucf(X,N). The number of errors is then estimated as N * Ucf(X,N). Starting at the leaves, we can then propagate this estimated error up the tree by equating the error of a subtree with the sum of the estimated errors of its children. Now, one can traverse the tree,

compare the number of estimated errors for the subtrees to the estimated error of replacing them by a leaf or branch, and make the replacement where predicted error is reduced. This approach invariably increases training set error but creates a more compacted tree, possibly with better predictive power. In our experiments, pruning reduced tree size from the unpruned tree, with no reduction in accuracy on test cases and some small increases.

A.3 Trees to Rules

One other function available in C4.5 that proves useful for our classification of original inputs and repeat corrections is the facility to translate from trees to rules. Rules can be easier to read than possibly large decision trees and may also illuminate some avenues for generalization. At the most direct level, rules can be written to summarize the tree by setting the antecedent of the rule to the conjunction of all tests on the path from the root to a given leaf. The consequent of the rule is simply the leaf label. C4.5 then attempts to simplify and generalize rules by deleting one or more conditions from the antecedent. The test for removing an antecedent uses a metric similar to the pessimistic error estimate used in the tree pruning procedure. One can build a contingency table describing the effects of deleting a condition from a rule that labels instances as some class C.

	Class C	Not C
Satisfies candidate condition	Y1	E1
Does not satisfy candidate condition	Y2	E2

The pessimistic error estimate for the original rule is computed as Ucf(E1,Y1+E1) and that for the new rule would be Ucf(E1+E2, E1+E2+Y1+Y2). If the error rate estimate for the new, simpler rule is no worse than that for the original. the new rule is retained. To remove multiple antecedents, C4.5 takes a greedy approach, always deleting the condition leading to the lowest error estimate. Next the system pursues an MDL approach to find the best set of rules to cover each class by trading off number of rules against the number of classification errors for that rule set. When these options are too numerous to test each case, a simulated annealing approach is used to explore the space of subsets, accepting any rule that decreases the length of encoding and those that increases the encoding only with some probability. The class whose rules cause the fewest false positives orders its rules before those causing more. The default class is chosen as the majority class for those instances not covered by any rule.

A.4 Issues in recognizing corrections

There are two important interactions between the functionality of decision tree classifiers, and C4.5 in specific, and the task of identifying spoken corrections. The first issue is that of the general shape of the decision regions defined by the classifier. Since

decision trees split instance sets through a sequence of tests on single attributes, they define hyper-rectangular decision regions. Now in the statistical analysis of acoustic features, we observed significant *proportional* increases in measures such as duration and pause, with large overlap in absolute ranges of values. This difference suggests the use of normalized or proportional measures, such as speaking rate or speaking rate divided by total duration or pause a proportion of total duration. In our experiments, these normalized measures proved useful, improving over absolute duration measures that were most useful only for the shortest utterances.

In addition, while the trends for increase in duration, pause, and pitch variability and decrease in pitch minimum are highly significant, they are not absolutely uniform. Some users buck the trends and as [Oviatt *et al.*, 1996] observe, in high error rate conditions adaptations become less marked over time. As a result, some of the features of corrections take on a probabilistic character. This variability suggests that fine-grained fitting will lead to overfitting. While pruning provides some help in relieving this problem, C4.5 provides an additional mechanism for limiting this type of overfitting. Specifically, the system allows the user to restrict the granularity of classification by requiring a minimum number of instances to be tested for splitting a decision node. In our experiments we achieve our best classification results when we restrict our tests to require a minimum of either cases for smaller data sets (176 instances) and a minimum of 10-20 cases for larger data sets (606 instances).

Adjusting features to make them more compatible with the rectangular decision regions defined by decision trees and limiting overfitting by constraining test size where noise, variability, or imperfect rectangularity interfere, improve the performance of decision tree classification for identifying spoken corrections.

Appendix B

Statistical Tests

This appendix describes common statistical tests used in the analysis sections of this thesis. It is simply intended as refresher for anyone whose statistics has gotten a bit rusty.

B.1 Analysis of Variance

Analysis of Variance (ANOVA) is a statistical hypothesis testing procedure. It is used to measure differences between population means to determine whether they differ. ANOVA can be used to compare two or more population means in one or more independent variables. In these respects, it is more general than the t-test discussed later in this appendix.

Analysis of variance works by comparing the variation within populations to the variance between populations. Differences are relevant when they are greater between than within populations. One first computes the variance between groups, or treatments, as the mean squared error between groups, or MS_b. Then one computes the variance within groups, or mean squared error, MS_w. Finally one compares these two measures by computing the F-ratio, that is simply $\frac{MS_b}{MS_w}$. This computation is often presented as an ANOVA Table as below:

Source	Sum-of-Squares	DF	MS	F
Group	$SS_b = \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$	k-1	$MS_b = \frac{SS_b}{\text{k-1}}$	$\frac{MS_b}{MS_w}$
Error	$SS_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$	N-k	$MS_w = \frac{SS_w}{\text{N-k}}$	
Total	$\sum_{i=1}^{k} \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$			

The null hypothesis, that the variances within and between groups is the same, leads to an F-ratio of 1. Since variance is always positive, F-ratio is always positive. The critical value of the F-ratio distribution depends on the number of degrees of freedom for the numerator and denominator. The number of degrees of freedom between groups, numerator, is 1 less than the number of columns (k). The number of degrees of freedom with groups is the number of observations (N) minus the number

of groups (k), or N - k. If the F-ratio exceeds the critical value for these degrees of freedom for some significance level (p), the null hypothesis is rejected and the populations are determined to be significantly different.

B.2 T-test

The most common form of test compares the means of two sample populations, with possibly different subjects and different numbers of subjects, in a commom experimental design. It assumes that the populations have normal distributions and similar variances. [1] There are two forms of hypothesis that can be tested about these means:

- Non-directional (or two-tailed).

  The means of the populations are diferent, if the null hypothesis can be rejected.

- Directional (or one-tailed).

  The mean of one population is greater than that of the other, if the null hypothesis can be rejected. The direction of the test is chosen by parameter setting.

Two-tailed tests are generally the most common.

To compare two populations, one computes a test statistic (t) of the following general form:

$$t = \frac{\text{sample statistic - population statistic}}{\text{estimated standard error}}$$

The sample statistic computes the difference between the sample means $(\overline{\chi}_1 - \overline{\chi}_2)$, while the population statistic computes the difference between the population means $(\overline{\mu}_1 - \overline{\mu}_2)$. The estimated standard error is often calculated using the combined (pooled) errors for the two populations.[2] This error measure is computed as follows:

$$s_p^2 = \frac{\text{Sum-of-squares-error1 - Sum-of-squares-error2}}{\text{df1 + df2}}$$

, where df1 + df2 = (n_1 -1) + (n_2 -1). Finally the estimated standard error is computed as $\sqrt{s_p^2/n_1 + s_p^2/n_2}$.

The resulting t-value is then compared to the associated critical value for the Student's t-distribution for some significance level, usually p ¡ 0.05 and the number of degrees of freedom. If the t-statistic exceeds that critical value, the null hypothesis can be rejected at level p.

[1] An alternative version of the test removes the latter assumption.

[2] This calculation differs when the variances are believed to be different.

Bibliography

[Allen *et al.*, 1987] Jonathan Allen, M Sharon Hunicutt, and Dennis Klatt et al. *From text to speech: the MITalk system.* Cambridge University Press, 1987.

[Bachenko and Fitzpatrick, 1991] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in Eng lish. *Computational Linguistics*, 16:155–170, 1991.

[Chung, 1997] Grace Chung. Hierarchical duration modelling for speech recognition. Master's thesis, Massachusetts Institute of Technology, 1997.

[Colton, 1995] D. Colton. Course manual for CSE 553 speech recognition laboratory. Technical Report CSLU-007-95, Center for Spoken Language Understanding, Oregon Graduate Institute, July 1995.

[Daly and Zue, 1996] Nancy Daly and Victor Zue. Statistical and linguistic analyses of f0 in read and spontaneous speec h. In *Proceedings of second international conference on spoken language p rocessing*, 1996.

[Davis and Hirschberg, 1988] James Davis and Julia Hirschberg. Assigning intonational features in synthesized spoken directions. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, 1988.

[Fernald *et al.*, 1989] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. De Boys son Bardies, and I. Fukui. A cross-language study of prosodic modifications in mothers' and father s' speech to preverbal infants. *Journal of Child Language*, 16:477–501, 1989.

[Heeman and Allen, 1994] P.A. Heeman and J. Allen. Detecting and correcting speech repairs. In *Proceedings of the ACL*, pages 295–302, New Mexico State University, Las Cruces, NM, 1994.

[Hirschberg and Litman, 1993] Julia Hirschberg and Diane Litman. Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, 19(3):501–530, 1993.

[Mann and Thompson, 1986] W.C. Mann and S.A. Thompson. Rhetorical structure theory: Description and construction of text structures. In Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligense, Psychology and Linguistics.* Kluwer Academic Publishers, 1986.

[Mingers, 1989] J. Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3(4):319–342, 1989.

[Mirgafiori *et al.*, 1995] N. Mirgafiori, E. Fosler, and N. Morgan. Fast speakers in large vocabulary continuous speech recognition: Analyses and antidotes,. In *Proceedings of Eurospeech 1995*, 1995.

[Nakatani and Hirschberg, 1994] C.H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustic Society of America*, 95(3):1603–1616, 1994.

[Nakatani *et al.*, 1995] Christine Nakatani, Julia Hirschberg, and Barbara Grosz. Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI-95 Spring Symposium in Palo Alto, CA on Empirical Methods in Discourse Interpretation*, pages 106–112, 1995.

[Nooteboom, 1997] Sieb Nooteboom. The prosody of speech: melody and rhythm. In William J. Hardcastle and John Laver, editors, *The Handbook of phonetic sciences.* Blackwell Publishers Ltd, 1997.

[Ostendorf *et al.*, 1996] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shribergand D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Proceedings of the International Conference on Spoken Language Processing*, 1996. supplementary paper.

[Oviatt *et al.*, 1994] S. Oviatt, P. Cohen, and M. Wang. Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication*, 15(3–4):283–300, 1994.

[Oviatt *et al.*, 1996] S.L. Oviatt, G. Levow, M. MacEarchern, and K. Kuhn. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 801–804, University of Delaware and A.I. duPont Instit., 1996.

[Oviatt *et al.*, 1998] S. Oviatt, M. MacEachern, and G. Levow. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24(2):87–110, 1998.

[Picheny *et al.*, 1986] M. Picheny, N. Durlach, and L. Braida. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29:434–446, 1986.

[Pierrehumbert, 1990] J. Pierrehumbert. The meaning of intonation in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press, 1990.

[Prevost, 1996] Scott Prevost. Modeling contrast in the generation and synthesis of spoken language. In *Proceedings of the ICSLP '96: the Fourth International Conference on Spoken Language Processing*, 1996.

[Quinlan, 1988] J. Quinlan. Decision trees and multi-valued attributes. In J. Hayes, D. Michie, and J. Richards, editors, *Machine Intelligence 11*, pages 305–318. Oxford, UK: Oxford University Press, 1988.

[Reichman, 1985] Rachel Reichman. *Getting Computers to talk like you and me.* MIT Press, 1985.

[Rudnicky and Hauptmann, 1990] Alexander Rudnicky and Alexander Hauptmann. Errors, repetition, and contrastive emphasis in speech recognition. 1990.

[Shriberg *et al.*, 1992] E. Shriberg, E. Wade, and P. Price. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfcation. In *Proceedings of the DARPA Speech and Language Technology Workshop*, pages 49–54. Morgan Kaufman Publishers: San Mateo, CA, 1992.

[Shriberg *et al.*, 1997] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Eurospeech '97*, 1997.

[Stifelman, 1993] Lisa Stifelman. Final Project, Computational Models of Discourse, CS 288, Harvard University, 1993.

[Sukkar *et al.*, 1996] Rafid Sukkar, Anand R. Setlur, Mazin Rahim, and Chin-Hui Lee. Utterance verification of keyword string using word-based minimum verification error (WB-MVE) training. In *Proceedings of ICASSP 96*, volume 1, pages 518–521, 1996.

[Swerts and Ostendorf, 1995] M. Swerts and M. Ostendorf. Discourse prosody in human-machine interactions. In *Proceedings of the ECSA Tutorial and Research Workshop on Spoken Dialog Systems - Theories and Applications*, June 1995.

['t Hart *et al.*, 1990] J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: an experimental phonetic approach to speech theory.* Cambridge University Press, 1990.

[Taylor *et al.*, 1996a] Paul Taylor, Hiroshi Shimodaira, Stephen Isard, Simon King, and Jaqueline Kowtko. Using prosodic information to constrain language models for spoken dialogue. In *Proceedings of 1996 International Symposium on Spoken Dialogue*, pages 129–132, 1996.

[Taylor *et al.*, 1996b] Paul Taylor, Hiroshi Shimodaira, Stephen Isard, Simon King, and Jaqueline Kowtko. Using prosodic information to constrain language models for spoken dialogue. In *Proceedings of ISSD 96: 1996 International Symposium on Spoken DIalogue*, pages 129–132, 1996.

[Taylor, 1995] Paul Taylor. The rise/fall/continuation model of intonation. *Speech Communication*, 15:169–186, 1995.

[Terken, 1997] J. Terken. Variation of accent prominence within the phrase: Models and spontaneous speech data. In Y. Sagisaka, W. Campbell, and N. Higuchi, editors, *Computing Prosody for Spontaneous Speech*, pages 95–116. Springer-Verlag, 1997.

[Yankelovich *et al.*, 1995] N. Yankelovich, G. Levow, and M. Marx. Designing SpeechActs: Issues in speech user interfaces. In *CHI '95 Conference on Human Factors in Computing Systems*, Denver, CO, May 1995.